DOCUMENT RESUME

ED 458 746 EC 308 692

AUTHOR	Elliot, Stephen N.; Braden, Jeffery P.; White, Jennifer L.
TITLE	Assessing One and All: Educational Accountability for Students with Disabilities.
INSTITUTION	Council for Exceptional Children, Arlington, VA.
ISBN	ISBN-0-86586-375-X
PUB DATE	2001-00-00
NOTE	197p.
AVAILABLE FROM	Council for Exceptional Children, 1110 North Glebe Rd., Arlington, VA 22201-5704 (Stock no. P5360: CEC members, $39.95; nonmembers, $49.95). Tel: 888-232-7733 (Toll Free); e-mail: service@cec.sped.org; Web site: http://www.cec.sped.org/.
PUB TYPE	Books (010) -- Guides - Non-Classroom (055)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	Academic Standards; *Accountability; Alternative Assessment; Case Studies; *Disabilities; *Educational Assessment; Educational Practices; Elementary Secondary Education; *Inclusive Schools; School Districts; Standardized Tests; *State Standards; *Student Evaluation; Testing
IDENTIFIERS	Testing Accommodations (Disabilities)

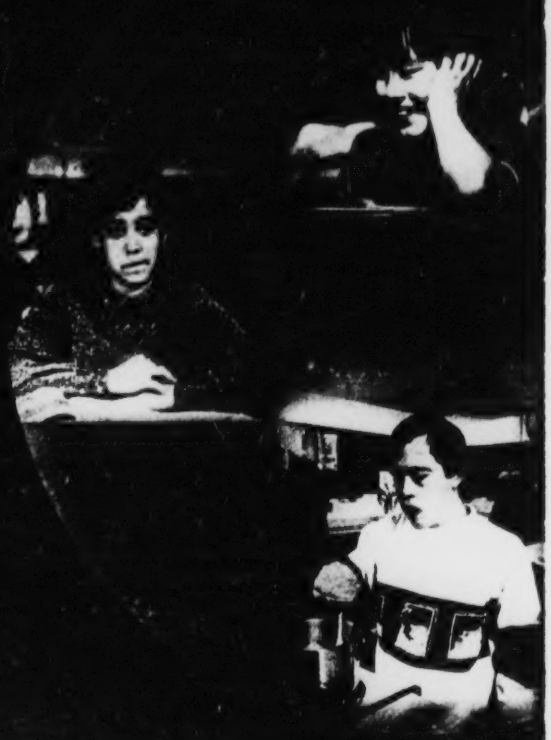
ABSTRACT

This book is about the assessment and inclusion of all students, including those with disabilities, in statewide and district assessment programs. It addresses aspects of assessment such as testing practices, test content, legal guidelines, technical aspects of tests, students' learning objectives, and instructional programs. Throughout the book, the cases of three students are featured: a fourth grader with reading difficulties, an eighth grader with learning disabilities, and an eleventh grader with Down syndrome. Students like these have often been excluded from state and district tests or, if tested, their scores may not have been reported. The five chapters address the following topics: (1) educational assessment today; (2) characteristics of good assessments; (3) understanding and using large-scale assessments; (4) inclusive assessment tactics such as testing accommodations and alternate assessments; and (5) best practices for inclusive assessment programs and educational accountability. Six appendices provide information on calculating the standard error of measurement, the Iowa Tests of Basic Skills, the Stanford Achievement Test (9th Edition), Terra Nova, and the Code of Fair Testing Practices in Education. A glossary of assessment and testing terms is also provided. (Contains 64 references.) (DB)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Assessing One And All

ED 458 746



PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Safer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Educational Accountability for Students with Disabilities

Stephen N. Elliott
Jeffery P. Braden
Jennifer L. White

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

EC 308692

BEST COPY AVAILABLE

Council for Exceptional Children

ASSESSING ONE AND ALL

Educational Accountability for Students with Disabilities

STEPHEN N. ELLIOTT
JEFFERY P. BRADEN
JENNIFER L. WHITE

Council for Exceptional Children

ISBN 0-86586-375-X

Copyright 2001 by Council for Exceptional Children, 1110 North Glebe Road, Suite 300, Arlington, Virginia 22201-5704

Stock No. P5360

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without prior written permission of the copyright owner.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

PREFACE v

ABOUT THE AUTHORS vii

CHAPTER ONE 1

Educational Assessment Today

CHAPTER TWO 15

Characteristics of Good Assessments

CHAPTER THREE 31

Understanding and Using Large-Scale Assessments

CHAPTER FOUR 67

Inclusive Assessment Tactics: Testing Accommodations and Alternate Assessments

CHAPTER FIVE 111

Best Practices for Inclusive Assessment Programs and Educational Accountability

APPENDIX A 119

Standards for Teacher Competence in Educational Assessment of Students

APPENDIX B 127

Calculating the Standard Error of Measurement

APPENDIX C 129

Iowa Tests of Basic Skills

APPENDIX D 141

Stanford Achievement Test (9th Edition)

APPENDIX E 157
TerraNova

APPENDIX F 175
Code of Fair Testing Practices in Education

GLOSSARY OF ASSESSMENT AND TESTING TERMS 181

REFERENCES 189

Preface

We wrote this book for teachers, administrators, school counselors, school psychologists, and parents who want to facilitate the meaningful participation of all students in assessments that are aligned with state academic standards. This book is about the assessment and the inclusion of all students in statewide and district assessment programs. In particular, it focuses on tactics for including students with disabilities in large-scale assessments to achieve a more complete picture of student learning and educational accountability.

As stressed throughout this book, assessing all students is an important and, at times, challenging undertaking that requires knowledge of testing practices, test content, legal guidelines, and technical aspects of tests, as well as a clear understanding of students' learning objectives and instructional programs. If educators across the United States are going to actualize the requirements of the Individuals with Disabilities Education Act (IDEA 1997) and the potential of standards-based education for all students, then all educators will need to be armed with a solid understanding of assessment fundamentals and details about assessment practices. This book is designed to help educators become familiar with a standards-based education framework and be knowledgeable of the general content of large-scale assessments used in a majority of states. In addition, the book provides detailed information on state testing guidelines, the valid use of testing accommodations, the valid use of alternate assessments, and how to communicate assessment results to educational stakeholders.

This book has been written to facilitate application of its content by featuring the cases of three students, Patrick, Tia, and Chris. Students like these three have often been excluded from state and district tests, or if tested, their scores may not have been reported. Such behavior has resulted in an incomplete picture of performance for these students and the schools they attend. Excluding students like Patrick, Tia, and Chris is no longer acceptable. Including these students in a meaningful way that results in valid assessments is an achievable goal for all educators. As a result of reading this book and talking with colleagues about assessment activities like those required by your state, you will be prepared to facilitate the meaningful participation of all students in statewide and district assessments.

The content and insights about assessment and students with disabilities in this book have been greatly influenced by our U.S. Department of Education funded research on testing accommodations and years of continuing professional development workshops on assessment with educators across the country. Our work with hundreds of Wisconsin's best teachers have taught us some valuable lessons about communicating information about technical aspects of assessment and fundamentals of large-scale assessments.

The need for a book like this one was recognized by leaders in Wisconsin's Department of Public Instruction in 1997. Thus, the two senior authors were commissioned to write a book specifically for Wisconsin teachers on *Educational Assessment and Accountability for All Students.* That book was published in 2000 and was highly acclaimed by teachers from Beloit to Rhinelander. Given its successful reception by administrators and teachers on the front lines, we decided to rewrite the book for a large audience. It is more comprehensive and case-focused, and consequently we think better than the original Wisconsin book. It includes new information on testing accommodations and alternate assessments, and covers the three most frequently used large-scale assessment instruments (i.e., Iowa Tests of Basic Skills, Stanford Achievement Test, and TerraNova) in the country. The book also is a companion to a Web-based course entitled "Assessing One and All," which is offered for continuing education or college course credits through the Council for Exceptional Children (www.cec.sped.org). This Web course brings the cases of Patrick, Tia, and Chris to life via video and audio vignettes and provides learners with links to their own state's academic standards, testing guidelines, assessment systems.

We have enjoyed writing this book and hope it helps you facilitate the meaningful participation of all students in state and district accountability systems, because all students count and should be counted.

Stephen N. Elliott
Jeffery P. Braden
Jennifer L. White
May 6, 2001

About the Authors

Stephen N. Elliott received his doctorate in Educational Psychology at Arizona State University in 1980. He is presently the Associate Director of the Wisconsin Center for Education Research and is a Professor in the School Psychology Program at the University of Wisconsin-Madison. He currently codirects four U.S. Department of Education grants concerning the use of testing accommodations with students with disabilities. Steve has been a productive scholar, authoring over 100 publications and 3 widely used behavior rating scales. Steve also has enjoyed the opportunity to consult with thousands of educators across the country about assessment issues and intervention services for students with disabilities. He is a former Editor of School Psychology Review and served 3 years on the National Academy of Sciences' Committee on Education Goals 2000 and Students with Disabilities.

Jeffery P. Braden received his doctorate in Educational Psychology at the University of California-Berkeley. He is a Professor of Educational Psychology at the University of Wisconsin-Madison. He codirects two U.S. Department of Education grants: one on assessment literacy and large-scale assessment and the second on the treatment utility of assessment practices with students with disabilities. Jeff has presented more than 100 papers at national and international meetings of psychologists and educators and has published over 100 articles, book chapters, and scholarly papers on assessment and students with special needs. Jeff is the lead author on the hypermedia course "Assessing One and All," which is the companion to this book.

Jennifer L. White is a doctoral student in Educational Psychology at the University of Wisconsin-Madison. She received her B.A. in Psychology and B.S.W. in Social Work there as well. Jennifer completed her practical training at the Waisman Center on Mental Retardation and Human Development, where she worked with children with disabilities and their families. She participated in the re-norming of the Woodcock-Johnson Tests of Cognitive Abilities III and currently serves as the senior project assistant on the U.S. Department of Education funded project entitled "Assessing One and All: An Internet Hypermedia Model for Professional Development."

CHAPTER ONE

Educational Assessment Today

High levels of student achievement are outcomes of schooling that virtually everybody values. Consequently, teachers and other educational professionals are expected to document student achievement and provide periodic summaries of educational progress to students, parents, and fellow educators. The process of documenting and reporting information about student achievement is dependent on good assessments and a method of communicating the results of these assessments so that they are meaningful. Collectively, good assessment and meaningful reports to the public about students' learning are the central ingredients of educational accountability.

Assessment is *not* a new activity for teachers. Most teachers engage in a wide range of assessment activities daily. For example, let's look into the classroom of Jackie Young, a fourth-grade teacher, with an eye toward the various assessment activities she undertakes during the course of a typical day.

■ Jackie arrived at school, as usual, 30 minutes before the first bus arrived. She readied her room for the day's activities by writing the work schedule on an overhead transparency, briefly organized her lesson notes, and then went to meet her students as they came streaming into the building at 8:15. During the course of the day, she:

- Recommended Josh spend extra time each night this week reviewing his multiplication facts.
- Called on Sandy twice even though she had not volunteered to answer questions about the social studies unit.
- Scored and assigned grades to her students' spelling tests.
- Referred Jason to the school psychologist for evaluation because of the persistent learning difficulties he was having in math and science.
- Stopped her planned English lesson halfway through the period to review the previous day's lesson because several students seemed confused.
- Assigned homework in math and social studies, but not English.
- Reviewed learning objectives for the forthcoming statewide assessment in mathematics and then made some minor adjustments in her lesson plans to include 2 days to do some sample test items.
- Held a lunchtime conference with the parents of a student with a disability to discuss the possible use of testing accommodations to facilitate his inclusion in the forthcoming state and district assessments.
- Gave a quiz in science covering two chapters and a field trip experience.

10

- Listened to oral book reports from half of her students and then provided them feedback about each of their presentations.
- Made notes to herself about some key words and important concepts in science that students were struggling with during a class discussion on rocks.
- Wrote three short essay questions and outlined model answers to each question in preparation for an end-of-unit test in social studies that she planned to give next week.

As illustrated by Jackie Young's vignette, **educational assessment is an information gathering and synthesizing process for the purpose of making decisions about students' learning and instructional needs.** Common assessment methods for most teachers include self-constructed tests or quizzes, interviews or oral questioning, classroom observations, behavior rating scales, classroom projects, and commercially published tests.

Today, with the advent of standards-based educational reforms and changes in laws concerning the assessment of all students, many educators involved in the assessment of student achievement need more advanced knowledge of assessment tools and practices. In particular, educators need more knowledge about the use and interpretation of standardized group achievement tests with all students because of the increased consequences associated with such tests in statewide assessment programs. Thus, this book on assessment of students has been written to advance teachers' understanding of assessment, in particular large-scale assessments, and ways to facilitate the inclusion of *all* students in assessments that are being used as the primary method for increasing educational accountability to the public.

Why Assess Students?

Teachers and parents obviously want students to learn and excel in school. Consequently, assessments are needed to determine whether students are learning and developing competencies that are needed for success later in life. Educators have observed that most students work harder and are more attentive when they think they are going to be held accountable for what they are studying. In other words, when students know that they will be assessed on the subject matter they are being taught, they tend to study harder and learn more. Thus, for some students, knowing that they are going to be assessed has important intentional and motivational consequences. It is widely recognized that tests can be a source of anxiety for some students and, concurrently, exciting opportunities to demonstrate what they know. Tests and assessments can also be sources of anxiety for educators. So why give tests and create statewide assessment systems? Tests play a major role in the lives of most students and teachers. They are used to:

- Measure student achievement.
- Evaluate students' acquisition and degree of mastery of important skills.
- Provide information to guide instructional practices.

- Evaluate the effectiveness of instructional practices.
- Monitor educational systems for public accountability.

Different types of tests and related assessment practices are needed to adequately achieve each of the various purposes just listed. Before getting too far into our examination of the various assessment practices educators use and the information resulting from these practices, it is important to have a good understanding of key assessment terms and fundamental assessment principles that should guide wise use of tests and assessment results.

Six Key Terms to Facilitate Communication

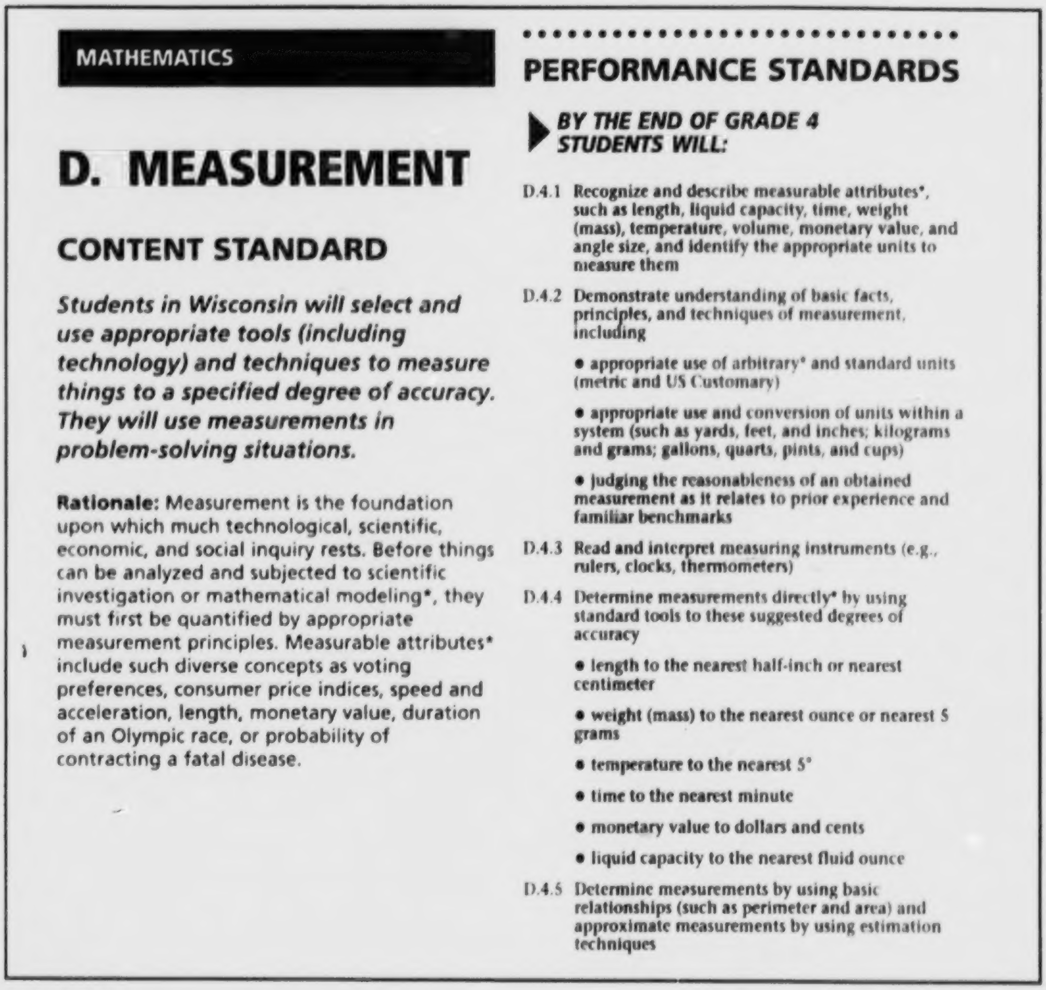
Communication about tests and educational assessment in an era of standards-based reform requires us to carefully define assessment, testing, and measurement and terms associated with standards, including content standards, performance standards, and proficiency standards.

By **assessment** we mean the process of gathering information about a student's abilities or behavior for the purpose of making decisions about the student. There are many tools or methods a teacher can use to assess a student, such as paper-and-pencil tests, rating scales or checklists, interviews, observations, and published tests. Thus, assessment is more than testing.

Testing is simply one procedure through which we obtain evidence about a student's learning or behavior. Teacher-constructed tests, as well as commercially published tests, have played and will continue to play a major role in the education of students. Such tests are assumed to provide reliable and valid means to measure students' progress. A test is a sample of behavior. It tells us something—not everything—about some class or type of behavior. Well-designed tests provide representative samples of knowledge or behavior.

To **measure** means to quantify or to place a number on a student's performance. Not all performances demonstrating learning can or need to be quantified (for example, art or musical exhibitions). The science of measurement in itself includes many important concepts—validity, reliability, standard scores—for teachers and others responsible for assessing students.

Educational assessment today is occurring within a context of educational change commonly referred to as *standards-based reform.* Every state has embarked upon some form of standards-based reform. Three types of standards are central to states' reform efforts. The first type is **content** or **academic standards**. These are general statements that describe *what students should understand and be able to do* in various content areas, such as English, language arts, mathematics, science, and social studies. Subsumed within each content standard are **performance standards**, which are defined as specific statements of expected knowledge and skills necessary to meet a content standard requirement at a particular grade level. Thus, performance standards indicate *how students can show what they understand and can do* (see Figure 1.1). Finally, **proficiency standards** are descriptive categories that describe the degree to which performance standards have been attained. In most states, there are four levels of proficiency used to describe *how well* a student has done on a test that is designed to measure most of the state's content standards (see Figure 1.2).



MATHEMATICS

D. MEASUREMENT

CONTENT STANDARD

Students in Wisconsin will select and use appropriate tools (including technology) and techniques to measure things to a specified degree of accuracy. They will use measurements in problem-solving situations.

Rationale: Measurement is the foundation upon which much technological, scientific, economic, and social inquiry rests. Before things can be analyzed and subjected to scientific investigation or mathematical modeling*, they must first be quantified by appropriate measurement principles. Measurable attributes* include such diverse concepts as voting preferences, consumer price indices, speed and acceleration, length, monetary value, duration of an Olympic race, or probability of contracting a fatal disease.

PERFORMANCE STANDARDS

BY THE END OF GRADE 4 STUDENTS WILL:

D.4.1 **Recognize and describe measurable attributes*, such as length, liquid capacity, time, weight (mass), temperature, volume, monetary value, and angle size, and identify the appropriate units to measure them**

D.4.2 **Demonstrate understanding of basic facts, principles, and techniques of measurement, including**

- **appropriate use of arbitrary* and standard units (metric and US Customary)**
- **appropriate use and conversion of units within a system (such as yards, feet, and inches; kilograms and grams; gallons, quarts, pints, and cups)**
- **judging the reasonableness of an obtained measurement as it relates to prior experience and familiar benchmarks**

D.4.3 **Read and interpret measuring instruments (e.g., rulers, clocks, thermometers)**

D.4.4 **Determine measurements directly* by using standard tools to these suggested degrees of accuracy**

- **length to the nearest half-inch or nearest centimeter**
- **weight (mass) to the nearest ounce or nearest 5 grams**
- **temperature to the nearest 5°**
- **time to the nearest minute**
- **monetary value to dollars and cents**
- **liquid capacity to the nearest fluid ounce**

D.4.5 **Determine measurements by using basic relationships (such as perimeter and area) and approximate measurements by using estimation techniques**

FIGURE 1.1
Sample Mathematics Content and Performance Indicators from Wisconsin's Model Academic Standards

Note. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 4. Copyright February 2000 Wisconsin Department of Public Instruction.

Much more will be said about educational assessment within a standards framework as you progress through this book. Thus, it is important that you have a good understanding of the six key terms that were just presented before reading further. In addition, it is important to keep children in mind when thinking about assessment. Three students for you to think about are Patrick, Tia, and Chris. These students have some wonderful abilities, but each also has some difficulties that interfere with learning. Read more about these three students in the box on page 6, because we will revisit their case studies periodically throughout this book.

BEST COPY AVAILABLE

Advanced	Proficient	Basic	Minimal Performance
Distinguished in the content area. Academic achievement is beyond mastery. Test score provides evidence of in-depth understanding in the academic content area tested.	Competent in the content area. Academic achievement includes mastery of the important knowledge and skills. Test score shows evidence of skills necessary for progress in the academic content area tested.	Somewhat competent in content area. Academic achievement includes mastery of most important knowledge and skills. Test score shows evidence of at least one major flaw in understanding the academic area tested.	Limited in the content area. Test score shows evidence of major misconceptions or gaps in knowledge and skills basic to progress in the academic content area tested.

FIGURE 1.2
General Proficiency Levels Used to Describe Student's Performance on the Statewide Knowledge and Concepts Examinations

Note. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 5. Copyright February 2000 Wisconsin Department of Public Instruction.

Principles to Guide Assessment

In many ways, large-scale assessment is a puzzling activity to many teachers. Historically, such assessments have not been aligned with standards, have used different tests about every 3 years, and have not been associated with any significant consequences. Times are changing, and periodic large-scale assessments are becoming an important part of educational accountability. In fact, many states are now beginning to develop formal decision rules for determining how and when a student will participate in large-scale assessments. Therefore, it may be useful to keep the following fundamental assessment principles in mind when you are discussing or using achievement tests to evaluate your students.

Principle 1: Standards First, Then Testing

When states and school districts set out to reform their educational systems, it is important that a logical sequence of events be followed. First, goals for each educational system should be set. Second, content standards need to be adopted that specify what children should know and be able to achieve. Third, curricula need to be adopted and instructional materials selected to help teachers help their students meet the standards. Finally, assessments should be developed to measure students' progress toward meeting the standards. In other words, "assessments should follow, not lead, the movement to reform our schools. . . . Only then can we build and use new tests that accurately measure students progress toward meeting standards" (Kean, 1998, p. 2).

Many of the desired skills and much of the information that educators value today are part of a state's *content and performance standards* that have been developed in the areas of reading, mathematics, language arts, writing, science,

Case Study Introductions

Patrick

■ Patrick is a 9-year-old 4th-grader who has difficulty reading. Patrick is a friendly and outgoing child but has always seemed a bit immature for his age. He has poor work habits and frequently loses his homework or forgets to do it.

Patrick began kindergarten on time at age 5. From early on, Patrick's teachers have had to work closely with him to keep him on task. He has always been easily distracted and works best in a highly structured environment. During the 2nd grade, Patrick moved with his mother from Arizona to a large urban school district in Florida. Soon after, Patrick's teachers noticed he was lagging behind the other students in reading achievement. Patrick was considered for retention, but this idea was decided against. Patrick's mother felt he was just having difficulty adjusting to a new school and insisted he be promoted to the 3rd grade. However, Patrick's reading difficulties only grew worse, and his classroom behavior began to deteriorate. Now as a 4th-grader he is experiencing more difficulty.

Tia

■ Tia is an 8th-grader who is classified as learning disabled. Her instructional reading level is fifth grade, but she receives all her instruction in regular classes with some support from a consulting special education teacher. She has good listening and memory skills, and is a highly motivated student.

Tia began kindergarten at age 6½ years. Both the school and Tia's parents felt it was in her best interest to wait a year before entering kindergarten. When she finally entered school, she struggled academically and socially. Tia was failing most subjects by the end of 2nd grade. Although her teacher considered retaining Tia, she instead referred her to determine whether she was eligible for special services. The IEP team reviewed test scores, class work, and other information, and decided Tia had a learning disability related to reading and language arts. She receives ongoing support from a consulting special education teacher in a mainstreamed classroom. For the last 5 years, Tia has received academic and social skills instruction and many classroom-based instructional accommodations. Tia's peer interactions and academic performance have improved dramatically.

Chris

■ Chris is a 17-year-old boy in the 11th grade. Chris was diagnosed at birth with Down syndrome. He works well with teachers and aides and has academic skills typical of a 2nd- or 3rd-grade student. He has some difficulty attending, but has been taking medication to improve his attention.

Chris has received a variety of supplemental educational supports through the years. Chris experienced multiple problems at birth and had to have open-heart surgery to correct an atrial septal defect. Chris has very sensitive hearing and will often cover his ears and hide his head when around loud noises. Chris's parents have been highly involved in his IEP planning from the start of his education and have been very supportive of the school's attempts to include Chris in the regular 11th-grade curriculum to the greatest extent possible. However, Chris receives most of his core instruction in a special education resource room, which he shares with 10 other students. He has two aides in addition to a resource teacher to assist him. Through his "Employability Class," Chris has begun a job helping to clean the lunchroom at various times throughout the day. Chris has performed very well in his job and enjoys it a great deal.

? Have students like Patrick, Tia, and Chris been successfully included in the large-scale assessments in your school district? If so, how? If not, how come? What could you do to facilitate their meaningful participation?

and social studies. Recent surveys have found that these standards vary widely from state to state. Much of this variation is due to differences in how states collect and make use of assessment data. However, use of results from tests that validly assess what all students know and can do in these content areas is a major component of a common *accountability system* for students receiving instruction in either a regular education or a special education classroom. Information about all students' educational performance lies at the core of any educational accountability system. Still, many of our current assessment systems do not account for every student within our public school system. As a result, our nation's understanding about how all student are achieving and how all schools are doing may be distorted and incomplete (National Center for Educational Outcomes [NCEO], 2000). Only with public reporting on these performances can policy makers and educators make informed decisions to improve education for all students. At this time, results of students' performances on achievement tests have become the most frequently used indicator for accountability purposes. Thus, involving all students in assessment systems is an important aspect of an inclusive education and is essential to educational accountability.

Principle 2: Tests Measure Educational Achievement; They Don't Create It

The central purpose of any test is to provide accurate and reliable information, not to drive educational reform. Some people have suggested that tests alone can create higher levels of educational achievement, but it is important to realize that new assessment systems cannot cure ailing education systems. Tests do not create better students. Rather, good teachers and good schools do.

Meaningful information resulting from tests, however, can help teachers do their jobs better. From a teacher's perspective, the primary purpose of assessment is to gather information about students' performances to make decisions about how and where the students should be instructed. Therefore, to the degree that teachers are knowledgeable about assessment, they increase the likelihood of making good decisions about the students in their classrooms. In essence, effective teaching boils down to good instruction, good assessment, and using each to do the other better (Witt, Elliott, Daly, Gresham, & Kramer, 1998).

Principle 3: No Single Test Does Everything; Thus, It Is Important to Use Multiple Measures and Repeated Measurements

Most educators realize that no single test can serve all the possible purposes for testing. A variety of tests or multiple measures are necessary to provide educators with a comprehensive view of what students know and can do. This should not be surprising given the array of learning expectations we have for students—we want them to be able to read, write, communicate orally, use technology, do research, calculate, conduct experiments, and understand and solve social problems. Some of these skills or competencies could be meaningfully assessed with a group-administered paper-and-pencil test that requires brief answers, while others would require more individualized assessments with direct observations by a teacher and the production of a product or detailed report. In light of this fact, the National Center on Educational Outcomes (NCEO) has recommended that states develop guidelines specifying how students with disabilities can be assessed in multiple settings using a variety of methods. Just as it is important to assess student performance in a variety of ways, it is sound practice to assess important skills or competencies at least twice to gain confidence in the assessment results.

Principle 4: Valid and Reliable Test Scores Are Important

For assessment results to be useful, the subject matter examined should be similar to what has been emphasized during instruction and students' responses must be measured and scored accurately. In the words of testing experts, an assessment must be *valid* and *reliable.* Tests that are used to make important educational decisions must meet rigorous technical standards for producing accurate and valid information.

The concepts of test score validity and reliability are quite abstract for most people and seemingly important only to the experts who construct tests. And yet almost every student we have ever worked with will express concerns about a test that doesn't appear to measure what he or she has been taught or results in inconsistent scores for two or more students who have produced very similar responses. Thus, students care about the quality of tests and the meaning of the scores that result from a test even if they don't understand the technical concepts of reliability and validity. Most educators and parents also care about the quality of tests, especially if important educational decisions such as promotion or graduation are based on such tests. However, these issues become extremely complex when applied to real-life cases. "How should the scores from alternate assessments be reported?" "Are scores obtained from out-of-level tests valid?" The answers to these questions are far from clear, and they continue to be the subject of considerable debate in the research literature. Consequently, we will be saying quite a bit about the concepts of reliability and validity in Chapter 2, especially in the context of inclusive assessment practices.

High Standards for All Students

You have probably read about or heard colleagues speak about high standards for all students, and you have no doubt wondered, Is this possible? Few educational movements have been so clearly identified by a single rallying cry as the standards-based reforms now dominating the nation's education policy agenda (*Education Week*, 2001; McDonnell, McLaughlin, & Morison, 1997). Central to the standards-based reform efforts is the belief that setting clear and high academic standards and expecting schools to teach and students to learn according to those standards can serve as a potent lever to improve overall educational quality. Four common elements seem to characterize this reform across the country. First, there is a focus on student achievement as the primary measure of school success. Second, there is an emphasis on challenging academic standards that specify the knowledge and skills students should acquire and the levels at which they should demonstrate mastery of that knowledge. Third, there is a desire to extend the standards to all students, including those for whom learning expectations have been traditionally low. Fourth, and one of the main concerns of this book, there is a heavy reliance on achievement testing to spur change and to monitor the reform's impact. Consequently, personnel in departments of education or public instruction across the United States have developed frameworks for educational standards, state assessments, and accountability systems.

Concurrent with the standards-based education reform efforts, there have been changes in federal law concerning students with disabilities and their involvement in all state and districtwide assessment programs. For example, the 1997 revisions of the Individuals with Disabilities Education Act (IDEA '97) now require states to establish goals for the performance of students with disabilities that are consistent, to the maximum extent appropriate, with other goals and standards for general education students established by the state. As such, states are now required to include children with disabilities in general statewide and districtwide assessment programs, with accommodations as necessary, or provide alternative assessment options for children unable to fully participate in large-scale assessment programs. States are also required to include the performance of students with disabilities in their official accountability reports. Thus, the goals of most standards-based reforms are to (a) specify in the form of academic and performance standards the knowledge and skills that all students will be expected to demonstrate at selected times during their education; (b) encourage educators to align their curriculum and instruction so as to facilitate students' opportunities to acquire the knowledge and skills competencies; (c) develop or purchase valid tests or other methods for assessing the extent to which all students achieve these knowledge and skills competencies; and (d) communicate annually with the public, using proficiency standards, to report how well students are performing with respect to identified knowledge and skills competencies. These are challenging goals, but not unrealistic.

Perhaps one of the most significant challenges for all of us in education is to establish high academic standards and document the results of all students' education against these standards across statewide or districtwide assessment systems. A particularly vexing part of this challenge is the meaningful participation of students with disabilities in one accountability system with all other students. Given that a significant number of students with disabilities and limited English proficiency (LEP) historically have been excluded or exempted from large-scale assessments, substantial efforts will be needed to achieve an accountability system that truly includes all students. For example, participation rates for students during the past several years in statewide assessments have ranged from a low of 33% to a high of 97% (Thurlow, Nelson, Teelucksingh, & Ysseldyke, 2000). Many of the students who did not participate were students with disabilities or with limited English proficiency.

For More Information on
Participation Rates of Students with Disabilities, see:

http://www.coled.umn.edu/NCEO

There are several possible reasons for the lower-than-desired participation rates of students with disabilities in our statewide assessments. These include:

- A perception that the tests are not relevant.
- A desire to "protect" these students from another frustrating testing experience.
- A concern that these students will lower the school's mean score in each content area.
- The fact that some parents do not want their son or daughter spending time taking a test that they don't understand or value.
- The belief that guidelines for administering a standardized achievement test prohibit, or at least limit, what can be changed without jeopardizing the validity of the resulting test score. Many educators have been admonished "Don't mess with the test," and so they are confused about what can and cannot be changed with a test.

If educators and other educational stakeholders who aspire to high standards for all students are to have a meaningful picture of how well students are learning and applying valued content knowledge and skills, all students need to be assessed periodically. The absence of students with disabilities from our statewide and districtwide assessment will result in (a) unrepresentative mean scores and norm distributions, (b) reinforcing beliefs that students with disabilities cannot do challenging work, and (c) undermining inclusion efforts for many students who can benefit from the same instruction as their peers without disabilities.

Testing students, making decisions about including students with disabilities in assessment programs, and implementing assessments so they are valid requires teachers' active involvement and can be challenging activities. As we noted earlier, some teachers may find large-scale assessments a bit puzzling. This is an understandable state of mind because there are a number of pieces to the "accountability puzzle" (see Figure 1.3) and there are some new legal requirements concerning students with disabilities. We have already introduced many of the pieces of the accountability puzzle, and in fact, have written this entire book around the key topics highlighted by this puzzle metaphor. Thus, at this time it is enough to simply familiarize yourself with the nine topics identified in the puzzle. However, over the course of reading this book, you will learn more about how these pieces of the accountability puzzle fit together and result in a big assessment picture.

For More Information on

IDEA, see:

http://www.ideapractices.org

Teachers Have Standards Too: Professional Roles and Responsibilities for High-Quality Student Assessments

We have only begun an examination of assessment of student achievement, and yet it should be evident that teachers must be knowledgeable assessment agents, capable of using a variety of techniques to describe students' learning

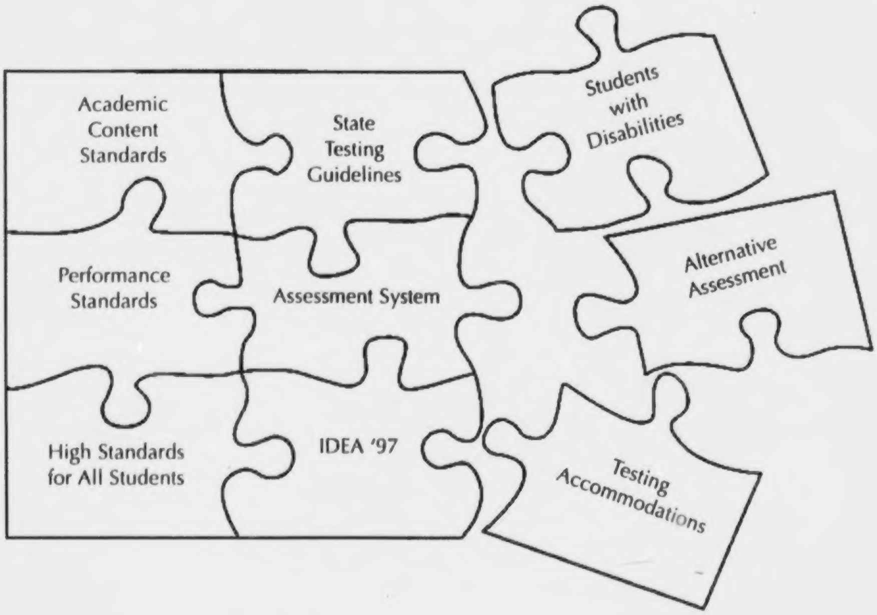

Academic Content Standards
State Testing Guidelines
Students with Disabilities
Performance Standards
Assessment System
Alternative Assessment
High Standards for All Students
IDEA '97
Testing Accommodations

FIGURE 1.3
Pieces of the Accountability Puzzle

Note. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 7. Copyright February 2000 Wisconsin Department of Public Instruction.

and to communicate with students, parents, and others about such learning. Accordingly, the American Federation of Teachers believes that "assessment competencies are an essential part of teaching and that good teaching cannot exist without good student assessment" (1990, p. 1). As a result of these beliefs, educators representing the American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association wrote a set of seven standards for teacher competence in student assessment. A brief listing of these standards follows (see Appendix A for a complete copy of *Standards for Teacher Competence in Educational Assessment of Students*):

> *Standard 1:* Teachers should be skilled in *choosing* assessment methods appropriate for instructional decisions.
>
> *Standard 2:* Teachers should be skilled in *developing* assessment methods appropriate for instructional decisions.
>
> *Standard 3:* Teachers should be skilled in *administering, scoring,* and *interpreting the results* of both externally produced and teacher-produced assessment methods.
>
> *Standard 4:* Teachers should be skilled in using *assessment results* when making decisions about individual students, planning teaching, developing curriculum, and improving schools.

Standard 5: Teachers should be skilled in *developing valid pupil grading procedures* that use pupil assessments.

Standard 6: Teachers should be skilled in *communicating assessment results* to students, parents, other lay audiences, and other educators.

Standard 7: Teachers should be skilled in *recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.*

The enactment of these standards for competencies in educational assessment requires a range of activities by teachers prior to instruction, during instruction, and after instruction. For example, assessment activities prior to instruction involve teachers' (a) clarifying and articulating the performance outcomes expected of students, (b) understanding students' motivations and creating connections between what is taught and tested and the students' world outside of school, and (c) planning instruction for individuals and groups of students that is aligned with what will be tested. Assessment-related activities occurring during instruction involve (a) monitoring student progress toward instructional goals, (b) identifying gains and difficulties students are experiencing in learning and performing, (c) adjusting instruction to better meet the learning needs of students, (d) giving contingent, specific praise and feedback, and (e) judging the extent to which students have attained instructional outcomes. Finally, the assessment-related activities occurring after instruction that involve teachers include (a) communicating strengths and weaknesses based on assessment results to students and parents; (b) recording and reporting assessment results for school-level analysis, evaluation, and decisionmaking; (c) analyzing assessment information before and during instruction to understand each student's progress and to inform future instructional planning; and (d) evaluating the effectiveness of instruction and related curriculum materials

It is important that special educators be well versed on student assessment practices as well. Special educators are often in a position in which they have to explain test results to concerned parents, develop individualized support plans, and make difficult decisions regarding educational placements. All of these activities require special educators to know how to accurately interpret and apply test results to complex cases. In recognition of this fact, the Council for Exceptional Children has included "assessment, diagnosis, and evaluation" among its "common core" of basic knowledge and skills essential to *all* special educators (1998, p. 19). These standards expand upon those adopted by the American Federation of Teachers and are now considered to be the minimal entry-level knowledge necessary to teach children with disabilities effectively. These standards include:

Standard 1: Special educators should be knowledgeable about the *basic terminology used in assessment.*

Standard 2: Special educators should be knowledgeable about the *ethical concerns related to assessment.*

Standard 3: Special educators should be knowledgeable about the *legal provisions, regulations, and guidelines regarding assessment of individuals.*

Standard 4: Special educators should be knowledgeable about the *typical procedures used for screening, prereferral, referral, and classification.*

Standard 5: Special educators should be knowledgeable about the *appropriate application and interpretation of scores, including grade scores versus standard score, percentile ranks, age/grade equivalents, and stanines.*

Standard 6: Special educators should be knowledgeable about the *appropriate use and limitations of each type of assessment instrument.*

Standard 7: Special educators should be knowledgeable about the *incorporation of strategies that consider the influence of diversity on assessment, eligibility, programming, and placement of individuals with exceptional learning needs.*

Standard 8: Special educators should be knowledgeable about the *relationship between assessment and placement decisions.*

Standard 9: Special educators should be knowledgeable about the *methods for monitoring progress of individuals with exceptional learning needs.*

To close this section on teachers' roles in assessment, we want to highlight a review study by Robert Hoge and Theodore Coladarci (1989) concerning research on the match between teacher-based assessments of student achievement levels and objective measures of student learning. As a rationale for their work, they noted that (a) many decisions about students are influenced by teachers' judgments of the students' academic functioning and (b) historically there seems to be a widespread assumption that teachers generally are poor judges of the academic abilities of their students.

Hoge and Coladarci identified 16 studies that were methodologically sound and featured a comparison between teachers' judgments of their students' academic performance and the students' actual performance on individualized achievement tests. They found generally high levels of agreement between teachers' judgmental measures and the standardized achievement test scores. The range of correlations was from a low of .28 to a high of .92, with the median being .65. (*Note:* A perfect correlation would be 1.00.) The median correlation certainly exceeds the validity coefficients typically reported for psychological tests.

In a recent replication of this research on the accuracy of teacher judgments, Demaray and Elliott (1998) found that teachers accurately predicted 79% of the items that a diverse sample of students actually completed on a standardized achievement test of reading and mathematics. The teachers in this study were virtually equally adept at predicting the achievement of students with high ability and students with below average ability. Collectively, the research on teachers' ability to judge the academic functioning of students has an important practical implication: Teachers, in general, can provide valid performance judgments of their students. This result is comforting, and it shouldn't be surprising given the number of hours that teachers have to observe their students' performances. The results, however, don't mean that tests are unnecessary, as some teachers who have heard about this research suggest. To meet the information needs of many educational stakeholders, we will continue to need periodic achievement test results for all students, as well as teacher judgments.

Assessment Is Communication!

We started this book with a dictionary-like definition of assessment—that is, assessment is an information gathering and synthesizing process for the purpose of making decisions about students' learning and instructional needs. We have stressed throughout this chapter that communication is a central part of, and perhaps the primary reason for doing, an assessment. In education we want to communicate how well students are learning to a wide array of people including students themselves, parents, administrators, legislators, and fellow teachers. If we are going to be successful in our communication efforts, teachers must have a strong command of assessment knowledge. Without this knowledge, the communications with the public and our fellow educators about student learning in the context of widely held standards will be far less meaningful and less effective. In summary, think of assessment as a communication activity rich with feedback and opportunities to tell a story about student achievement and educational effectiveness.

CHAPTER TWO

Characteristics of Good Assessments

Good educational assessments yield good scores. Educational assessments come in many forms, including traditional multiple-choice tests, observations of students' work samples, and extended responses or performances. As emphasized in Chapter 1, they serve a variety of purposes. But regardless of the type of assessment or its purpose, all good assessments should possess the characteristics of *validity, reliability,* and *usability.* For many readers, these are familiar terms commonly associated with tests and testing. And yet their meaning is not well understood. Many readers will automatically assume we are about to present advanced statistics and some esoteric measurement concepts that have little to do with their teaching lives. This is not the case. Instead, this chapter focuses on practical concepts that are central to assessing students and using the results of any assessment with confidence. In this short but important chapter, we define and discuss three characteristics of good assessments and provide some guidelines for using this information when you select or construct your own assessments.

As an educator, you occasionally will have to explain the significance of an assessment, especially a large-scale assessment mandated by your school district or state. The involvement of students with disabilities in such assessments likely will stimulate even more inquiries about the validity and reliability of the resulting scores if testing accommodations or an alternate assessment have been used. Therefore, knowledge about validity, reliability, and usability are important in the delivery of effective assessment services.

Before examining these three key assessment concepts, let's establish how we typically use achievement tests and the resulting test scores. Basically, an achievement test is given once or possibly twice a year to a group of students with the intent of providing a score for each student that is indicative of his or her knowledge or ability in a given subject matter area. *The resulting test scores are useful or good to the extent that the test (a) measures what the students have been studying in their classes and (b) the resulting scores are accurate.* To the extent that the test measures subject matter content that is different from what students have been studying, students' test scores become less meaningful as indicators of their achievement and less useful in guiding teachers' future instructional efforts. Likewise, if the students' answers do not result in a test score that can be determined consistently and accurately, teachers' confidence in the score is lessened.

In summary, we tend to find achievement tests useful when they are representative of what students have been taught and when they yield consistent, accurate scores. When these conditions have been met, we are more comfort-

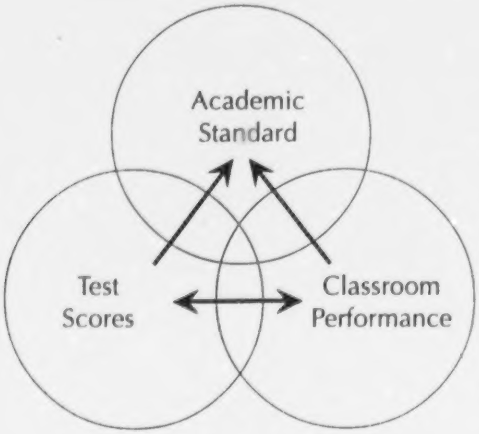


FIGURE 2.1
Illustration of the Desired Relationships Among Academic Standards, Scores on Test, and Students' Classroom Performance

Note. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 12. Copyright February 2000 Wisconsin Department of Public Instruction.

able or confident making inferences from the resulting test scores about students' classroom performances. When academic standards (like some of those in state content standards) have influenced classroom instruction, then it is logical to also consider a possible relationship between students' test scores and such standards. That is, it is reasonable to use test scores in a subject matter area as evidence of the degree to which students have acquired the knowledge and skills specified in content standards. The next chapter, in which we focus on widely used large-scale assessments, will examine further the relationship or alignment among standards, tests, and instruction. For now, examine Figure 2.1 to get a picture of the connections and associated inferences between a student's test scores and his or her classroom performances in mathematics, as well as the relationship between both of these and academic standards in mathematics. The inferred connections among these elements of the education system may be logical, but they are only meaningful if the resulting test scores are valid. If you are to make sound inferences about students' achievement, it is critical that tests like that used in your state's assessment system yield valid test scores.

Validity

When you test a student in basic mathematics, you are testing a *sample* of that student's mathematical knowledge and skills. From the resulting test score, you make an inference about the student's ability to add, subtract, and so forth. Your inference depends on the truthfulness or meaning of the test—its validity. *Validity refers to the adequacy and appropriateness of the interpretations made from*

Case Reflections and Good Assessments

Patrick

■ Recall that Patrick is a 4th-grader in Florida and is experiencing some difficulties with reading. Tests that require reading material written at a 3rd- and 4th-grade level may prove challenging for him to take. Thus, his results on the Florida statewide assessment, which includes tests in reading, mathematics, social studies, and science, all may be influenced by his reading difficulties. Consequently, Patrick's teachers and parents must decide whether his test results will be reliable and valid indicators of his knowledge and skills.

Tia

■ Remember that Tia is an 8th-grade student identified with learning disabilities and is experiencing some difficulties with reading fluency. In Wisconsin, where she lives, the statewide test is given in 4th, 8th, and 10th grades and covers the areas of language arts, mathematics, social studies, and science. Each of these areas requires a significant amount of reading in order to understand and answer test items. Tia's reading difficulties may well influence her resulting performances in each of these subject matter areas. Consequently, Tia's teachers and parents will need to decide whether her test scores in mathematics, social studies, and science can be considered valid without some accommodations.

Chris

■ Chris, as you recall, is an 11th-grader identified with Down syndrome, who receives all of his instruction in a special education classroom. The content of his daily curriculum is quite different from that of students in the general curriculum. Given this fact, it must be decided whether the Iowa Tests of Basic Skills, the statewide test given to the vast majority of students, would provide meaningful results about Chris's knowledge and skills. If not, how can his knowledge and skills be measured so that reliable and valid scores are achieved?

? Is it possible to reliably and validly assess students with disabilities on the same test as students without disabilities? What is the likely effect of a student's reading difficulties or disabilities on his or her test scores? How would you know if a test score was invalid?

BEST COPY AVAILABLE

assessments with regard to a particular use. Of all the essential characteristics of a good test, none surpasses validity. If a test is not valid for the purpose used, it has little or no value. For example, if a test designed to measure academic achievement in geography or history has questions that are phrased in difficult language, it probably does not test geography or history as much as it does reading. The test does not do a good job of measuring what it primarily claims to measure. Validity is specific. That is, a test may be valid for one purpose and not the others. For example, administering a spelling test for the purpose of determining a student's achievement in grammar is very likely to be invalid.

Traditionally, test developers have talked about three major kinds of validity: content validity, criterion-related validity, and construct validity. A test has *content validity* if it adequately samples knowledge and skills that have been the goal of instruction. Does the test adequately represent the material that was taught? Testing a minor portion of a unit on *Hamlet* after stressing the unity of the total play greatly diminishes content validity. Determining whether a test has content validity is somewhat subjective. It usually is established when subject matter experts and experienced teachers agree that the content covered is a representative sample of the knowledge and skills in the tested domain of knowledge and skills.

A test is said to have *criterion-related validity* if its results parallel some other external criteria. Thus, test results are similar or not similar to another sample of a student's behavior (i.e., some other criterion for comparison). If students do well on a standardized reading test that measures many aspects of reading, they likewise should do well in completing and understanding geography and history assignments. Some people refer to this type of validity as *predictive validity,* because a score from one assessment is being used to make predictions about a performance on another assessment that occurs later.

A test has *construct validity* when the particular knowledge domain or behavior said to be measured is actually measured. For example, a teacher may claim that his or her test measures application of mathematical concepts and not just mathematical computations. Therefore, a review of the test should reveal that large portions of the items require students to apply results of mathematical computations using mathematical concepts correctly. To further substantiate that the test measures the application of mathematical concepts, one could look for agreement between the test results and other evidence from students' classroom activities and work samples. Construct validity is a complex issue and increasingly is coming to refer to the entire body of information about what a test measures. As you can see in our example of the assessment of mathematical applications, decisions about construct validity require information about the content of the test and the degree to which the test results relate to other measures of the same construct.

It makes no sense to prepare or select a test designed to measure something other than what has been taught if you want the results to affect instruction and provide information about student learning. As an example, we don't measure a student's height using a bathroom scale. Therefore, teachers and others should work hard to ensure that a test measures what it is designed to measure. When it does, we say it has good construct validity.

Factors Influencing Validity

Numerous factors can make assessment results invalid for their intended use. Some are obvious and avoidable. For example, no teacher would think of measuring knowledge of mathematics with a social studies assessment. Nor would it be logical to measure problem-solving skills in fourth-grade mathematics with an assessment designed for eighth graders. In both instances, the assessments would yield invalid results.

Some of the factors that influence validity are subtle. A careful examination of test items or assessment tasks will indicate whether the assessment instrument appears to measure the subject matter content and the mental functions that the teacher is interested in measuring. However, several factors may prevent or interfere with the test items or assessment tasks functioning as intended. When this happens, the validity of the interpretations of the assessment results is diminished. Linn and Gronlund (1995) identified a list of 10 factors inherent in a test or the assessment itself that can interfere with valid results. These factors are listed and briefly described in Figure 2.2.

Factors involved in the administration and scoring of a test also may affect the validity of test results. With classroom assessments, factors such as insufficient time, unfair aid to individual students, cheating, and inaccurate scoring can lower validity. When using published tests, failure to follow the standard directions and time limits, giving students unauthorized assistance, and unreliable scoring contribute to lowering the validity of the results. Factors associated with changes in the administration of a test and the validity of the resulting scores are central to the use of testing accommodations with students with disabilities. Consequently, many teachers who administer assessments to all students will be confronted with decisions concerning the validity of the results for students with disabilities who received accommodations in the administration of a particular test or assessment. The appropriate use of testing accommodations should result in increasing the validity of the inferences made from a student's test score. Much more will be said about the issue of test score validity and the use of testing accommodations in Chapter 4.

Factors associated with students' responses to test items or assessment tasks can also affect the validity of the results. As Linn and Gronlund (1995) observed, some students may be bothered by emotional problems that interfere with their test performance. Others may be frightened or anxious in a testing situation and unable to respond as they would in daily classroom situations. Still others may not be motivated to put forth their best effort. We are also aware that some students with disabilities may need accommodations in the response format or method for reporting answers to test items. These and other factors that change students' responses to an assessment can distort results and consequently lower validity if the assessment is not implemented with care.

Evidence of Validity

Evidence of the validity of a score on a test or an assessment instrument generally takes two forms: (a) how the test or assessment instrument "behaves" given the content covered, and (b) the effects of using the test or assessment

1. **Unclear directions.** Directions that do not clearly indicate to the student how to respond to the tasks and how to record the responses will tend to reduce validity.
2. **Reading vocabulary and sentence structure too difficult.** Vocabulary and sentence structure that are too complicated for the students taking the assessment will result in the assessment's measuring reading comprehension and aspects of intelligence, which will distort the meaning of the assessment results.
3. **Ambiguity.** Ambiguous statements in assessment tasks contribute to misinterpretations and confusion. Ambiguity sometimes confuses the better students more than it does the poor students.
4. **Inadequate time limits.** Time limits that do not provide students with enough time to consider the tasks and provide thoughtful responses can reduce the validity of interpretations of results. Rather than measuring what a student knows about a topic or is able to do given adequate time, the assessment may become a measure of the speed with which the student can respond. For some content (e.g., a typing test), speed may be important. However, most assessments of achievement should minimize the effects of speed on student performance.
5. **Inappropriate level of difficulty of the test items.** In norm-referenced tests, items that are too easy or too difficult will not provide reliable discrimination among students and will therefore lower validity. In criterion-referenced tests, the failure to match the difficulty specified by the learning outcome will lower validity.
6. **Poorly constructed test items.** Test items that unintentionally provide clues to the answer will tend to measure the students' alertness in detecting clues as well as mastery of skills or knowledge the test is intended to measure.
7. **Test items inappropriate for the outcomes being measured.** Attempting to measure understanding, thinking skills, and other complex types of achievement with test forms that are appropriate only for measuring factual knowledge will invalidate the results.
8. **Test too short.** A test is only a sample of the many questions that might be asked. If a test is too short to provide a representative sample of the performance we are interested in, its validity will suffer accordingly.
9. **Improper arrangement of items.** Test items are typically arranged in order of difficulty, with the easiest items first. Placing difficult items early in the test may cause students to spend too much time on these and prevent them from reaching items they could easily answer. Improper arrangement may also influence validity by having a detrimental effect on student motivation. This influence is likely to be strongest with young students.
10. **Identifiable pattern of answers.** Placing correct answers in some systematic pattern (e.g., T, T, F, F or A, B, C, D, A, B, C, D) will enable students to guess the answers to some items more easily; and this will lower validity.

FIGURE 2.2
Inherent Factors That Influence Validity

Note. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 14. Copyright February 2000 Wisconsin Department of Public Instruction.

instrument. Questions commonly asked about a test's "behavior" concern its relation to other measures of a similar construct, its ability to predict future performances, and its coverage of a content domain. Questions about the use of a test typically focus on the test's abilities to reliably differentiate individuals into groups and to guide teachers' instructional actions with regard to the subject matter covered by the test. Some questions also arise about unintended uses of a test or an assessment instrument. For example: Does use of the instrument result in discriminatory practices against various groups of individuals? Is the test used to evaluate others, such as parents or teachers, whom it does not directly assess? These questions concern a relatively new area of validity referred to as *consequential aspects of validity* (Green, 1998; Messick, 1989), which is discussed in greater detail in Chapter 5 of this book.

Criteria for evaluating the validity of tests and related assessment instruments have been written about extensively (Linn & Gronlund, 1995; Witt et al., 1998). A joint committee of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education recently revised their comprehensive list of standards for tests that stresses the importance of construct validity and describes a variety of forms of evidence indicative of a valid test. These revised *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999) include valuable information for educators involved in testing diverse groups of students, including both students with disabilities and students with limited English proficiency.

Key Aspects of Validity

Many test users and consumers of test-based information struggle with the relatively abstract concept of validity and its importance to the meaningful use of tests or assessments. Be assured, however, that it is the single most important characteristic of good assessment information and must be understood by all test users. Keep in mind the following key aspects of validity noted by leading measurement experts (Airasian, 1994; Linn & Gronlund, 1995):

- Validity is concerned with the general question "To what extent will this assessment information or test score help me make appropriate decisions?"
- Validity refers to the decisions that are made from assessment information, not the assessment approach or test itself. It is not appropriate to say "This assessment information is valid" unless you also say for what decisions or groups it is valid. Keep in mind that assessment information valid for one decision or group of students is not necessarily valid for others.
- Validity is a matter of degree; it does not exist on an all-or-nothing basis. Think of assessment validity in terms of categories: highly valid, moderately valid, and invalid.
- Validity involves an overall evaluative judgment. It requires an evaluation of the degree to which interpretations and uses of assessment results are justified by supporting evidence. Educators also must consider assessment results in terms of the consequences of those interpretations and uses.

Although validity may be the most important characteristic of a good assessment, it is by no means the only characteristic you should understand. Consumers of test results also want the results to be reliable, so let's examine what reliability means with respect to test scores.

Reliability

A test is reliable to the extent that a student's scores are nearly the same on repeated measurements. It is characterized as reliable if it yields consistent scores. Suppose, for example, that a teacher has just given an achievement test to her students. How similar would the students' scores have been had she assessed them yesterday, or next week, or in a couple of months? How would the students' scores have differed if she had selected a different sample of tasks to test? How much would the scores have differed if another person scored the test? These are the types of questions with which reliability is concerned.

Remember, assessment results merely provide a limited measure of performance obtained at one point in time. Some error always exists in any test or assessment since fluctuations in human behavior are not totally controllable, and the test itself may contain possibilities of error. As errors in measurement increase, the reliability of a test decreases. Unless an assessment can be shown to be reasonably consistent over different occasions, different raters, or with different samples of tasks from the same subject matter, we can have little confidence in the results.

Carefully note the relationship and distinction between reliability (i.e., consistency) and validity (i.e., meaningfulness). A valid test must be reliable, but a reliable test need not be valid. In other words, *reliability is a necessary but not sufficient condition for validity.* For example, giving an algebra test to first or second graders will produce consistent results, but the results are not meaningful for 6-year-olds. Thus, the test would be reliable, but not valid.

Reliability can be described numerically and is primarily statistical, but please don't let that discourage you from learning more about it. It is important if you are going to be involved in using test results, and essential if you are ever going to design and conduct an alternate assessment for a student with a severe disability. The logical analysis of an assessment will provide little evidence concerning the reliability of the resulting scores. To evaluate the consistency of scores assigned by different raters, two or more raters must score the same set of student performances. Similarly, an evaluation of the consistency of scores obtained in response to different forms of a test or different collections of performance-based assessment tasks requires the administration of both test forms or collections of tasks to an appropriate group of students. Whether the focus is on interrater consistency or consistency across forms or collections of tasks, consistency may be expressed in terms of shifts in the relative standing of students in the group or in terms of the amount of variation to be expected in a student's score. We report consistency in the case of interrater judgments or across forms of a test by means of a *correlation coefficient.* In the case of the expected amount of variation in a given student's test score, however, we report consistency by means of a statistic called the *standard error of measurement.* Both of these methods of expressing reliability are widely used, and educators responsible for communicating the results of assessments should understand them.

Correlations can range between +1.0 and –1.0, where +1.0 indicates perfect agreement between the magnitudes of the scores for the same individual. The case of a test–retest approach to reliability is illustrated in Figure 2.3. Given that most teachers do not repeatedly administer a test, alternative methods of estimating the reliability of a test, such as internal consistency, must be used. The latter method uses a slightly different formula for calculating a reliability coefficient (referred to as *coefficient alpha*). Regardless of the method for quantifying the reliability of a test, most experienced users of teacher-constructed tests consider reliability coefficients in the +.80 or higher range to be essential. Many published tests have reliability coefficients in the +.90 range.

Test-Retest Reliability of a Kindergarten Screening Test

	Number of Answers Correct	
Student (N)	**Test (X)**	**Retest (Y)**
1	9	10
2	7	6
3	5	1
4	3	5
5	1	3

$$r_{xy} = \frac{N\,\Sigma XY - (\Sigma X)\,(\Sigma Y)}{\sqrt{\left[N\,\Sigma X^2 - (\Sigma X)^2\right]\left[N\,\Sigma Y^2 - (\Sigma Y)^2\right]}}$$

$$r = \frac{150}{\sqrt{(200)(230)}}$$

$$r = .70$$

FIGURE 2.3
Example of How to Calculate Test–Retest Reliability

Note. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 16. Copyright February 2000 Wisconsin Department of Public Instruction.

The *standard error of measurement* (SEM) is an estimate of the variation expected in a student's score if the student is given the same test over and over. The amount of variation in the scores is directly related to the reliability of the assessment procedures. Low reliability is indicated by large variations in the resulting scores, and high reliability by little variation in the scores.

It is impractical to administer the same test to a student repeatedly. Fortunately, however, it is possible to estimate the amount of variation in the resulting scores. This estimate of the variation in scores is the SEM. The calculation of the SEM for a test is beyond the scope of this book (but if you are really interested in how it is done, see Appendix B), and besides, most manuals of published tests provide specific standard errors of measurement. All you need to do is be able to apply your knowledge of SEMs when interpreting a student's test results. It is a wise practice to interpret a test score as a band of scores (which most people call a *confidence band*) rather than as a specific score. Chapter 3 will provide more details about SEMs and confidence bands.

Factors Influencing Reliability

Although teachers seldom find it possible or useful to calculate reliability coefficients or SEMs, they should be cognizant of factors that can influence assessment results. Two such factors are the number of items or tasks on a test and the objectivity of the scoring of the items or tasks.

In general, the larger the number of tasks on an assessment, the higher the reliability will be, because a longer assessment will provide a better sample of the knowledge and skills being measured. In addition, the scores are less likely to be distorted by chance factors.

Objectivity of an assessment refers to the degree to which equally competent scorers obtain the same results for the same students. Most of the published tests educators use are high in objectivity and are often scored by machines or highly trained scorers. In general, tests featuring selected-response (i.e., multiple-choice) items can be scored more reliably than constructed response items. Concerns about the reliability of scores, frequently voiced as issues of bias or fairness, often have been used to argue against the use of complex constructed response type tasks on achievement tests. However, with training it is possible to get highly reliable scores for written essays or performance tasks with multiple parts.

Key Aspects of Reliability

We can conclude our examination of reliability, then, by saying that unless a test is reasonably consistent on different occasions or with different samples of the same behavior, we can have very little confidence in its results. A variety of factors, some concerning the student taking the test and others inherent in the test's design and content, can affect the reliability of a test. Student characteristics affecting a test's reliability include guessing, test anxiety, and practice in answering items like those on the test (Witt et al., 1998). Test characteristics that can influence reliability include a test's length (longer tests are generally more reliable), homogeneity or similarity of items (more homogeneous tests are usually more reliable), and time allotted (speed tests are typically more reliable than unbound tests).

In conclusion, when considering the reliability of any test or assessment process, keep the following points in mind:

- Reliability refers to the stability or consistency of assessment information, not the appropriateness of the assessment information collected.
- Reliability is a matter of degree; it does not exist on an all-or-none basis. It is expressed in terms of degree: high, moderate, or low.
- Reliability is a necessary, but not sufficient, condition for validity. An assessment that provides inconsistent results cannot be relied upon to provide useful information. If important educational decisions are to be made from a test, the resulting score(s) must be highly reliable.

Usability

So far we have argued that good assessments should measure what they say they measure and that the measurements must be consistent—that is, good assessments are valid and reliable. Good assessments also must be useful. This may seem like an obvious point, but educators should not overlook it when designing or selecting an assessment, particularly when the assessment involves a large number of children. For example, in many statewide assessment systems more than 300,000 students are eligible to take a test each year. Thus, issues concerning ease of administration, interpretation and application, time required to administer the test, and cost should be weighed against alternative ways of getting the same information and the resulting consequences.

Unlike the concepts of validity and reliability, there is no general set of guidelines or statistical indices used to determine the usability of a test or an assessment program. A wide array of variables influence decisions about usability, and often they are the subject of debate.

One of the issues most hotly debated in assessment for educational accountability is how useful test results are for teaching and learning. When students as a whole do poorly on a test, there are two possible reasons for their poor scores: either the test is a poor measure of student learning, or the test accurately reflects the fact that students did not learn. Whether or not a test is a poor measure (and therefore not usable for making instructional decisions) is primarily determined by the concept of *alignment*—that is, whether the test is a good (i.e., reliable and valid) measure of the curriculum or standards students are to master. We will discuss the concept of alignment in greater detail in Chapter 3, because it is essential to the usability of large-scale assessments. However, if the test is aligned with the curriculum (i.e., what students are to master), then teachers can use assessment results to evaluate student learning—and their instruction. Good assessment results suggest that students learned and, by implication, that the teacher taught the subject matter effectively. Poor assessment results suggest that students did not learn and, by implication, that the teacher did not teach the subject matter effectively.

Many groups view the outcomes of accountability assessments as useful for evaluating schools, districts, and states. That is, schools or districts with high scores are viewed as offering a better education than schools or districts with low scores. The validity of this interpretation depends on many issues, some of which are not well supported by research (Haertel, 1999). However, results can be useful to teachers for making decisions about instruction. If students score well, teachers get useful feedback suggesting that their teaching methods are working. Conversely, if students score poorly, it suggests that teachers should change their instructional practices. We suggest a strategy for making decisions about instruction in Figure 2.4, Using Test Results for Instructional Decisionmaking. Note that we recommend that you verify the test results before you assume they are an accurate indication that students have not learned. This illustrates one of the fundamental principles of educational assessment—that you should verify results by using multiple methods of assessment.

Another key usability issue concerns how the results of an assessment are communicated. When results are stated in understandable terms to most consumers, but especially teachers, it increases the likelihood that they will facilitate teachers' instructional efforts and advance an understanding of their own abilities for students and their parents. An example of this is to report scores as proficiency levels or categories. Also, the specificity of results influences their usefulness. Knowing that 68% of students in a school district are proficient is not as useful as knowing that 82% of the students in your classroom have mastered basic understanding skills in reading, but only 33% of your students have mastered evaluation and extension of meaning in reading. Related to *how*

Q: Did Students Do Poorly on the Test?

No. Results point to adequate or superior performance.
Decision: Celebrate the success! Be sure to retain your instructional strategies; they are successful.

Yes. Results point to inadequate performance (i.e., low passing rates).
Decision: Go to next question.

Q: Does the Test Measure a Curricular Objective?

No. The objective is not part of the school or district curriculum for that grade level.
Decision: Ignore test results and return to step 1 for other objectives, or include the objective in the curriculum and continue as if the answer is "yes."

Yes. The objective is an appropriate expectation for that grade, and should be part of the student outcomes.
Decision: Go to next question.

Q: Is the Test Accurate?

No. A sample of students who failed test shows that they pass the test under other conditions, or have higher pass rates on similar objectives in a second test.
Decision: Either: (1) change the test, or (2) consider changing how children are prepared for test (see below).

Yes. A sample of children who failed test shows they fail it even when tested under other conditions.
Decision: Go to next question.

Q: Was the Objective Taught?

No. Careful examination of permanent products (e.g., texts, curricula, teaching activities) shows that the objective was not taught, or was taught inadequately (e.g., it was placed just before vacations, or near the end of the year).
Decision: Either (1) plan by the year, and have periodic (e.g., monthly or quarterly) review of annual teaching plans, or (2) consider team teaching or other mechanisms to ensure teacher coverage.

Yes. The objective was included in instructional materials and taught.
Decision: The students did not learn the material despite adequate exposure. Go to next question.

Q: Are there Other (More Effective) Ways to Teach the Objective?

Yes. All children are capable of learning!
Decision: Brainstorm reasons why instruction failed, and then develop alternatives to instruction (e.g., collaborative teaching, select alternative curriculum/materials, consult the research to identify high-strength instructional strategies and curricula).

FIGURE 2.4
Using Test Results for Instructional Decisionmaking

results are communicated is the issue of *when* results are communicated. For feedback of any kind to be useful, it must occur close in time to the performance of interest. Far too often, test results—particularly those from large-scale assessments—come months after the testing event occurred and with little time to focus on remediation efforts, and they may provide only large-group, general results for the fundamental subject matter areas.

Applying Knowledge of Good Assessments to Your Work

As emphasized in this chapter, good assessments are valid, reliable, and usable. Many educators have translated this "holy trinity" of measurement to mean that a test must measure what it says it measures and do so in a way that is practical and results in consistent scores. This is an acceptable translation, but perhaps a bit of an oversimplification of the judgments required of persons involved in using an assessment. Recall that validity is not an all-or-none characteristic of an assessment, but a matter of degree. Also remember that reliability is a necessary but not sufficient condition of validity. Ultimately, a statement about the validity of an assessment involves an evaluative judgment of the degree to which interpretations and uses of the assessment results (i.e., scores or proficiency statements) are justified.

To make decisions about the degree to which an assessment yields valid results, it is useful to ask the following four questions:

> ***The Content Question.*** How well does the sample or collection of assessment tasks *represent* the domain of tasks to be measured? For most teachers this question is answered by reviewing copies of tests and comparing the items to what they teach. The greater the similarity, the more confidence they have that the test measures what they value.
>
> ***The Test–Criterion Relationship Question.*** How well do students' performances on the assessment *predict* future performances or *estimate* current performances on some valued measure of the knowledge and skills other than the test itself? For most teachers, this question is answered by comparing the assessment results with another measure of performance, such as classroom tests or summary observations by the teacher. The greater the similarity between the test and teachers' other criterion of performance, the more confidence teachers have in the test scores.
>
> ***The Construct Question.*** How well can teachers interpret performance on the assessment as a meaningful measure of the knowledge and skills the assessment purports to measure? For most teachers, answers to this question will be out of reach, because it requires establishing the meaning of the assessment by experimentally determining what factors influence students' performances. Many educators will fall back on their review of the content and test–criterion relationships as evidence that the test measures a specific construct. Construct validation takes place primarily during the development of a test and is based on an accumulation of evidence from many sources. If you are using a published test or assessment program to measure a particular construct such as mathematical reasoning or reading comprehension, then you will find the necessary evidence on the con-

struct validity of the instrument included in a technical manual that accompanies the test.

The Consequences Question. How well does use of the assessment results accomplish the intended purposes of the assessment and avoid unintended effects? If an assessment is intended to contribute to improved student learning, the consequences question becomes deceivingly simple: "Does it?" In trying to answer this question, teachers typically pose many more questions. For example, "What impact does the assessment have on teaching? What are the possible negative, unintended consequences of the use of the assessment results?" As you can see, there is no short or easy answer to the consequences question. Nevertheless it is worthwhile to address it. In fact, it is often the first question many educators ask when confronted with a large-scale assessment program. We will revisit the topic of consequential validity in Chapter 3, after you have had a chance to learn more about the intended uses of large-scale achievement tests and the use of testing accommodations for students with disabilities.

Next to validity, reliability is the most important characteristic of a good assessment. Reliability provides the consistency that makes validity possible, and it indicates the degree to which various kinds of generalizations are reasonable. High reliability is essential when test results are going to be used to make final decisions that concern individual students and have lasting consequences. Under these conditions, the tests or assessments used should have a very small standard error of measurement and one should be able to readminister and rescore them to establish the consistency of the score(s), especially if a student's original score is below a critical cut-point. Lower reliability is tolerable when the test results are used to make reversible decisions of relatively minor importance and when the decision is confirmable by other data.

Finally, it is not enough to have tests or assessments that yield valid and reliable scores. The tests or assessments also must be usable. That is, persons with limited assessment training must be able to administer them, and the tests must be constructed to allow a wide range of students to participate in the assessment. Of course, time and costs are also important usability factors, as is the ease of interpretation. Ultimately, issues of usability influence validity; that is, if educators do not use an assessment as designed, they are unlikely to achieve the intended purpose of the assessment.

Many readers of this book will be working with students with disabilities and trying to facilitate their meaningful involvement in state and district assessment programs. As a result, they will find themselves having to make a number of decisions about the validity of assessment results. Specifically, when students need testing accommodations, teachers will be expected to select and implement accommodations that do not invalidate test results. The use of a testing accommodation, in fact, is intended to enhance the validity of the test score for the student with a disability. In addition, when a student cannot meaningfully participate in the regular assessment given to the majority of students, teachers and their fellow IEP team members will be responsible for conducting an alternate assessment. In many cases, teachers will play a major role in constructing these alternate assessments for an individual student. The alternate assessments, however, still will need to be valid and reliable. Consequently, knowl-

Case Applications and Good Assessments

Patrick

■ Patrick's state uses a commercially produced test that you will learn more about in the next chapter. Suffice it to say that the test has been developed to yield reliable and valid scores. However, the fact that Patrick is a poor reader and is not eligible to receive any testing accommodations in Florida suggests that his scores on tests of mathematics, science, and social studies may not be highly valid indicators of his true skills in these subject matter areas. As you recognize, significant reading difficulties can influence any student's performance on a test in which reading is needed to access and use information.

Tia

■ Tia's state also uses a commercially produced test that is well aligned with state academic standards. This test will be discussed in more detail in the next chapter, but it, too, is well developed and has significant evidence that it generally yields reliable and valid test scores. Tia's reading disability, if appropriately accommodated, should not have a negative effect on the validity of her test scores in mathematics, science, and social studies. Her performance on the reading test, however, cannot be accommodated, because reading is the skill that is being measured. Influencing her reading by using a reading accommodation would result in invalidating the reading test score. (In a handful of states, reading the reading test is allowed, but the resulting scores are reported as "nonstandardized." Tia lives in a state where this is not allowed.)

Chris

■ Chris's state uses a highly regarded commercially produced test that has substantial reliability and validity evidence. However, Chris will not be taking this test, because his curriculum focuses on functional skills. He will be taking an alternate assessment. Reliability and validity are still important issues to consider when reporting Chris's performances on Idaho's alternate assessment. Consequently, the educators conducting the alternate assessment will be responsible for documenting that the results are reliable and valid. Clearly, educators must really understand these technical concepts of good tests if they are going to conduct alternate assessments.

edge of the characteristics of a good assessment is critical to using test results and to facilitating the meaningful participation of all students in large-scale assessment programs.

In conclusion, issues pertaining to decisions about validity of test results start before a test is given, are ongoing after a test is completed, and are always relative to the stated purpose of the test. As you can see, the typical and seemingly straightforward question "Is the test valid?" requires some technical knowledge to answer and is actually worded inappropriately. Better questions, and ones you should be equipped to address, are: "Is the test a good test?" and "Does the test yield valid scores?"

CHAPTER THREE

Understanding and Using Large-Scale Assessments

This chapter will provide you with an understanding of large-scale assessment (LSA). Understanding LSA will help you do two things: First, you can better align curriculum, instruction, and assessment to improve outcomes for students. Second, knowledge of LSA content and results will help you decide whether, and how, to include students with disabilities in LSA.

This chapter has three sections and is supported by three appendixes. First, we will explain why LSA is important. Second, we will describe the types of results reported from LSA. Third, we will give you an opportunity to apply your knowledge of LSA to understanding sample outcomes. We conclude with some common questions and answers regarding LSA. We also provide three appendixes to help you understand the three most popular LSA tests used in schools, districts, and states. These appendixes describe: (a) the Iowa Tests of Basic Skills (published by Riverside; Appendix C), (b) the Stanford Achievement Test (published by Harcourt Brace; Appendix D), and (c) the TerraNova (published by CTB/McGraw-Hill; Appendix E).

Before you begin, reflect on LSA and the students in our three case studies: Patrick, Tia, and Chris. All of these students may be affected by LSA. This chapter will help you understand ways in which LSA affects their lives and will provide a better understanding of how it might affect the lives of the students you serve.

Why Have Large-Scale Assessment?

Although some educators embrace assessment, others view LSA as a necessary evil—or just plain evil. Educators' resistance to mandated assessment is understandable, because testing programs are often required by external agencies and may be used for many purposes that educators do not embrace, such as rating school districts or determining student promotion and graduation. However, there are two reasons why educators engage in assessment programs. The first reason is that you should; the second is that you must.

Why You Should Assess

Effective schools coordinate three features to enhance educational success: curriculum, instruction, and assessment (CIA). Schools that carefully align curriculum, instruction, and assessment enhance the performance of individual students. Schools that do not align curriculum, instruction, and assessment are less effective than schools with strong alignment (see Cotton, 1999, for a

40

Case Reflections on Large-Scale Assessment

Patrick

- Patrick's mother and teacher are concerned that he may not perform well on the upcoming Florida 4th-grade achievement test. Patrick's mother is worried that if he does not perform well and is retained, his problem behaviors will only escalate. He is not a strong reader, and he is doing average work at best in other areas.

Tia

- This year Tia's class is preparing for the upcoming Wisconsin 8th-grade exam. Although Tia would like to do well on the test, she is worried that she will perform poorly on the reading/language arts and writing sections of the test. Tia's teachers have always allowed her to use her notebook computer to check her grammar and spelling on written assignments, use books on tape, make recordings of class discussions, and get oral directions, and they often grant her extra time to complete work. Now, both Tia and her mother are concerned that if Tia fails a portion of the exam, she may not be promoted to high school next year.

Chris

- Chris's parents think he should be allowed to take the upcoming 11th-grade achievement test. They feel that it is his right to take the test just like any other student would. However, Chris's teacher is afraid he may not be able to handle the pressure of standardized testing. On more than one occasion, she has seen Chris become emotionally explosive when forced to work under pressure. Although she feels strongly that Chris should be included as much as possible, she doesn't feel that taking the 11th-grade exam at this time will do him, or any of the other students, any good.

? Why should these (or any other) students participate in large-scale assessments? What are the content and structure of the tests they will take? What kinds of results come from these tests? What are the consequences attached to these tests?

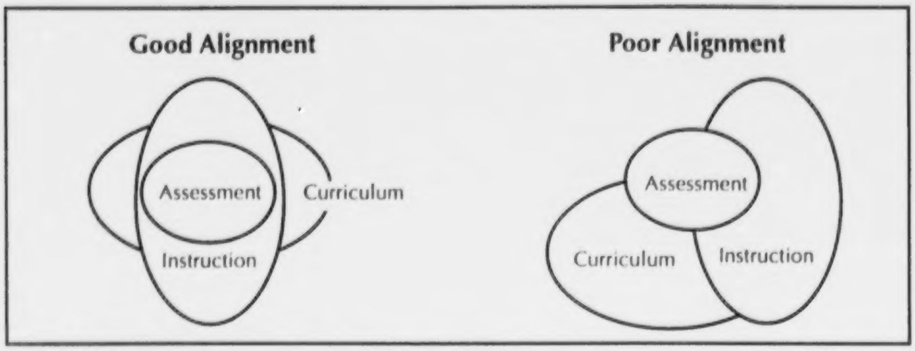


FIGURE 3.1
Curriculum, Instruction, and Assessment Alignment

Note. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 22. Copyright February 2000 Wisconsin Department of Public Instruction.

summary of how curriculum alignment enhances school effectiveness). When curriculum, instruction, and assessment are aligned, students understand what is expected, teachers understand what to teach, and schooling is more effective. Assessment of student progress is an essential catalyst for aligning curriculum and instruction. One might say good assessment functions as a "CIA agent"—that is, assessment stimulates alignment of curriculum and instruction to ensure student learning. Figure 3.1 illustrates appropriately and inappropriately aligned curricula.

There are at least two ways to align curriculum, instruction, and assessment. The first method is the heart of standards-based educational reform. In this method, states delineate curricula so that teachers, administrators, parents, and students understand curricular scope and intent. When curricula are clearly delineated, then teachers can select instructional practices to promote the outcomes specified in the curricula. The final step in this method of instructional alignment is assessment. That is, after educators specify the curriculum students are to master and provide instructional activities to promote mastery, they must assess students' performance on curricular objectives. This last step tells educators the degree to which they have been successful. It also informs students, parents, and the community at large of the effectiveness of schooling. State educational standards are intended to stimulate this top-down alignment process. That is, by telling the public, educators, and students the content students are to master at various stages of educational progress, the state intends to stimulate local school districts and educators to align their curricula and instructional practices to achieve state standards.

LSA measures how well schools do in helping students meet state standards; assessment stimulates accountability to ensure curriculum–instruction alignment. This process is endorsed by the American Federation of Teachers (Glidden, 1998) and is promoted as a means to improve student learning (Hammer, 1998; Novello, 1999), although some question the value of such top-down standards-based reforms (e.g., Barton, 1999; Taylor, 1994).

However, there is another way to align curriculum, instruction, and assessment. This method for alignment begins with assessment. That is, educators start by assessing student performance. Although beginning at the "end" of the CIA process appears illogical, it can be a powerful way for educators to take control of student learning. In fact, the most effective school reform typically begins with educators clarifying the outcomes they desire from students rather than beginning with curricula and teaching practices (Newmann, Marks, & Gamoran, 1995). Thus, educators can use the assessment of educational outcomes as a starting point, rather than an end point, for CIA alignment. Understanding the results of LSA can help educators achieve that alignment and, in turn, better educational outcomes for students. In this approach, educators clearly identify the outcomes of education—what they want students to know and do—and then align their assessments to measure these outcomes, their curriculum to reflect the outcomes, and their instruction to support the outcomes (see Newmann & Wehlage, 1995).

Why You Must Assess

Even if one does not accept the need for assessment as an essential ingredient in CIA alignment, most U.S. educators must formally assess student learning. Federal laws, such as the Improving America's Schools Act (IASA) and the Elementary and Secondary Education Act (ESEA), make federal funding contingent upon annual evidence of continuous school improvement—that is, continued improvement in student performance. These laws do not specify statewide testing, but all states have adopted statewide testing programs to meet the federal requirements for reporting annual student progress. State legislatures typically mandate that students in elementary, middle, and high school take state tests to meet federal requirements. Legislatures usually mandate that the results of these assessments be reported to the public. In many states and districts, additional consequences for students (e.g., promotion to the next grade, graduation from high school) may be attached to test results. Thus, educators must participate in LSA to meet state and district laws and regulations, and many students must participate to advance in or graduate from school (Heubert & Hauser, 1998).

Whereas federal and state mandates promote LSA for educational accountability purposes, other federal mandates (e.g., the Americans with Disabilities Act, the Individuals with Disabilities Education Act Amendments of 1997) require participation of all students in LSA whenever possible. The National Center for Educational Outcomes (1999) suggests that states should include 98% or more of all students (or about 85% of students with disabilities) in educational accountability programs such as LSA. Because of federal mandates, and concerns about fairness, states and districts direct educators to include all students with disabilities in LSA whenever possible. One goal of this chapter is to enhance your knowledge of LSA content so that you can make effective decisions about the inclusion of students with disabilities in LSA.

Myths and Realities and Pros and Cons of Large-Scale Assessment

LSA is a somewhat controversial practice. Some of the controversy surrounding LSA is deserved, because it is an imperfect tool for educational accountability. Other controversies surrounding LSA either fail to consider the positive aspects of LSA or derive from distortions or myths about LSA. Table 3.1 provides a list of some common myths and realities about LSA.

LSA is often criticized (and supported) on the basis of myth rather than reality. In fact, LSA is like most other educational practices in that it has some direct evidence of positive consequences, some direct evidence of negative consequences, and many claims (for and against) that are not clearly substantiated in the literature. Although many of the arguments in favor of LSA are not yet fully supported by research (Haertel, 1999), and some claims may be overstated (Linn, 2000), it is clear that LSA enjoys substantial political (Dorn, 1998) and popular support as a tool for educational reform. LSA also provides advantages and disadvantages for education. We list some of the pros and cons of LSA in Table 3.2 to help you evaluate what LSA does—and does not—do to promote education.

Large-Scale Assessment Structure and Content

Before explaining test results, it is a good idea to describe the content and structure of large-scale assessments. Most LSA systems strive to be fair. A fair large-scale assessment should provide (a) common tasks or tests (so all students get equally difficult tests); (b) information concerning what is on the test (so teachers know what to teach, students to know what to study, and parents know what their children are supposed to learn); and (c) information about how learning is evaluated (e.g., what kinds of tests or tasks will be used to measure students' knowledge).

Teachers generally view assessments customized to particular students and content as the most fair, because these assessments respond to the content demands and the students' unique characteristics. Consequently, teachers often view standardized, large-scale assessments as unfair, because LSA is not customized to students or the classroom. However, administrators, parents, and community members view customized assessments as having elements of unfairness, because some teachers may teach (and test) more demanding or less demanding curricula and may have different standards for judging student performance. Therefore, LSA, for accountability purposes, demands that all students, teachers, and parents have similar opportunities—and demands—for learning. In fact, by requiring a common assessment, standards-based accountability systems essentially require teachers, schools, and districts to provide similar opportunities to learn for all students. LSA is intended to increase instructional equity for all students, especially students who have historically received a different, often less rigorous education (e.g., ethnic minorities, students with disabilities). If students are excluded from LSA, it is easy to provide a less demanding education. The equity consequence of LSA can be put bluntly: "If students are not counted, they don't count."

Consequently, LSA demands common tasks for all students to ensure equity of opportunity to learn for all students. Although much of this book is devoted to helping you learn how to accommodate individual student needs within LSA, we firmly believe that inclusion and participation in LSA—using similarly challenging tasks, tests, and demands—is an essential step toward achieving educational equity for all students, especially students with disabilities.

TABLE 3.1
Myths and Realities of Large-Scale Assessment

Myth	*Reality*
• Testing takes too much time.	Students tested annually spend less than 1% of their school time taking tests. To put this in perspective, the typical elementary school student spends18 times more time in recess (i.e., 7% of the year, excluding lunch); the typical high school student spends 20–30 times more time in study hall (about 12% of the school year).
• Testing improves learning.	Just as you can't fatten cattle by weighing them, you can't improve students' performance by testing them. Testing helps achievement only to the degree that it improves instruction; testing in the absence of instructional changes does nothing to improve learning. However, accountability for student outcomes may induce instructional changes, which in turn may improve learning.
• Testing costs too much.	LSA typically consumes less than 5% of a state's education budget; when federal and local revenues are included in the resource pool, testing consumes less than 2% of annual educational expenditures. Businesses typically expend 5%–15% of their budgets in product and consumer assessment. Although some argue that testing costs outweigh benefits (e.g., Haney, Madaus, & Lyons, 1993), others argue tests benefits exceed costs (e.g., Phelps, 1996).
• Norm-referenced tests only assess rank order; not academic proficiency.	The tests used in LSA typically mirror national academic content standards in specified content areas. Scores have meaning for content mastery (criterion-referenced) and for relative mastery (norm-referenced) interpretations.
• To get high test scores you must teach test-taking skills or test-wiseness, not academic content.	The major factor in test scores is content knowledge, not test-wiseness. Teaching to standardized test-like tasks does not improve scores on standardized tests relative to group problem-solving instruction (Baxter et al., 1993). Also, coaching to improve test-wiseness does not improve scores to a large degree (Samson, 1985).
• Tests are sufficient measures of academic performance.	Tests are not sufficient; they must be supplemented by other methods for measuring academic performance and must be supplemented by other measures (e.g., teacher-made tests, portfolios, performance assessments, teacher judgments)—especially when making high-stakes decisions (American Educational Research Association, 2000).
• Tests are neutral measures of performance.	Tests sample only part of a state's educational standards; consequently, the standards included in the test are likely to receive more instructional attention than other standards, leading to excessive narrowing of the curriculum to which students are exposed.
• Tests damage students' self-esteem.	Although testing may be frustrating for students (especially poor performers), there is no evidence to suggest long-term, pervasive harm. Conversely, exclusion from testing may itself be harmful to self-esteem, by setting the student apart from peers and suggesting academic incapability. Studies of students with disabilities suggest that they are no more nor less frustrated by testing than nondisabled peers (Elliott & Kratochwill, 1998).
• Students with disabilities are tested too much.	Although it is true that students with disabilities are tested often and at length, these tests typically focus on eligibility for special services. Testing for planning and evaluating instruction is conducted far less frequently—and, many would argue, not often enough.

TABLE 3.2
Pros and Cons of Large-Scale Assessment

Pros	*Cons*
Cost. LSA is the least expensive way to provide a common measurement of academic performance for a large number of students. It is much less expensive than performance assessments, portfolios, and the like.	***Arbitrariness.*** LSA emphasizes selected-response items, which provides a less authentic measure of students' skills. Also, selected-response formats are poorly aligned with constructed response instructional assessments (e.g., projects, essays, portfolios).
Psychometric integrity. LSA has known (high) reliability, and its validity (e.g., alignment with state standards, relationship to other measures of achievement) is formally studied and usually acceptable.	***Limited learning value.*** Whereas projects, portfolios, and other constructed response assessments induce learning, students rarely learn new content or skills from LSA testing.
Accountability. LSA provides a common method by which to compare students, districts, etc., with respect to academic proficiency. Annual LSA also offers the opportunity to track student growth (i.e., how much students gain) via value-added accountability methods (see Meyer, 1996).	***Narrowing the curriculum.*** LSA may lead to teachers' spending more time teaching test-taking skills in the mistaken belief that test-wiseness substantially affects scores; also, educators tend to emphasize subjects and standards included on the test, consequently deemphasizing academic standards, and entire domains (e.g., music, art, physical fitness) are excluded from the test.
Focus on learning versus teaching. By providing LSA, states encourage students, parents, and educators to focus on student learning rather than focusing on instructional methods and materials.	***Teaching irrelevant material.*** Although there is little evidence to show that teaching test-taking skills has much effect on test scores, some teachers adopt test-preparation curricula in the hope of improving test scores.

Almost all states have standards in reading, English/language arts, mathematics, social studies, and science to inform teachers, students, parents, and the community what students in each state are expected to know and to do at various grade levels. Note that three states do not: Rhode Island has standards in three academic areas, Pennsylvania in two, and Iowa does not have statewide standards (*Education Week*, 2001). You can find information about your state or district educational standards on the Web, in publications, and from your principal or administrator. Moreover, most states publish specific information about how their LSA aligns with state standards—that is, how and how well their LSA measures state educational standards. This information is critical to understanding what will be on the test for all LSA stakeholders.

For More Information on

Your state standards, see:

www.achieve.org

The backbone of most state and district LSAs is a standardized test of achievement. All states except Nebraska have a multiple-choice test as part of their LSA; 38 states include short-answer items; and 46 states require an extended-response (i.e., essay) answer in English. However, only seven states require extended responses in subject areas other than English (e.g., Maine, New York), and only two states (Kentucky and Vermont) require portfolio methods to assess student learning (see *Education Week*, 2001).

Most states do not develop their own large-scale tests (i.e., multiple-choice and short-answer items). Typically, states negotiate with test publishers to produce a statewide or district test that matches the state or district standards. A state typically requests bids from companies to produce a test that meets the state standards. More than 40 states contract with one of three publishers to produce their state test: CTB/McGraw-Hill (publisher of the TerraNova), Harcourt Brace (publisher of the Stanford Achievement Test), or Riverside Publishing (publisher of the Iowa Tests of Basic Skills). Because these three tests so often serve as the backbone of state and district assessments, we have provided appendixes addressing their characteristics. Appendixes C, D, and E provide you with information about the typical number of items, reliability, validity, and format of each test as currently provided by the publisher. However, the characteristics of the test as offered by the publisher may vary from the version sold to a state. For example, New York and Indiana each purchase a slightly different version of the TerraNova from CTB/McGraw-Hill; California and Florida purchase somewhat different versions of the Stanford Achievement Test from Harcourt Brace, and so on. To find out which test your state or district uses, consult your state website or other resources.

For More Information on

Your state website, see:

http://www.ccsso.org/seamenu.html

Item content, or what the test tests, also is important to understanding LSA. The content of test items relates to academic objectives. For example, a language arts objective might be to "analyze text," which is shown by drawing conclusions, inferring relationships, and identifying theme and story elements; a mathematics objective might be "data analysis, statistics, and probability," which is shown by analyzing, interpreting, and evaluating data and applying concepts and processes of data analysis, statistics, and probability to real-world situations. Most LSA items used for state accountability systems are designed to assess specific academic objectives in each subject matter area. We provide examples of these academic objectives in Appendixes C, D, and E, which describe the most popular tests. However, you can find more information about which objectives are tested in your state or district by consulting the guides that accompany the test. For example, CTB/McGraw-Hill publishes *Teacher's Guide to TerraNova* (CTB/McGraw-Hill, 1997), which includes a thorough description of the academic objectives covered in each subject matter area in the test. Other publishers provide similar guides for their tests.

Appendixes C, D, and E provide examples of items from the popular tests. As you look at these examples, ask yourself the following three questions:

1. What kind of response does the item require from the student?

2. What academic objective or skill does the item require from the student?
3. What thinking skill does the item demand from the student?

By asking yourself these questions, you will better understand how LSA developers integrate item response formats, academic skills, and thinking skills into their assessment of students. Take time to explore test item content and format. You will need to thoroughly understand test content and response demands to make accurate decisions about how and when to include students with disabilities in LSA.

Understanding LSA Results

Types of Results

Tests evaluate student learning by the number and difficulty of the questions students answer correctly. However, reporting the number of correct answers to parents and students is not very useful. For example, saying your student got 30/40 items correct does not tell you much about how well your student did. If the test was very difficult, the 30/40 might represent an exceptionally good performance; if the test was exceptionally easy, 30/40 might be failing. Likewise, if the standard for accuracy is 50%, 30/40 is good; but if the standard is 90% accuracy, 30/40 is poor.

To understand a test score, you need to know two things: how the score compares to other students' scores and how the score compares to a given performance standard. Scores that tell how a student does relative to other students are called *norm-referenced* scores. Scores telling how a student does relative to a performance standard are called *criterion-referenced* scores. Neither type of score is sufficient to explain performance. Knowing a racer finished fifth in a 10K race, or knowing a salesperson sold $250,000 in products one month is only part of the story. You need to know the racer's time to fully understand whether the racer ran well or was just matched against weaker runners; likewise, $250,000 in sales may be good, average, or poor relative to other salespeople's totals. Both norm- and criterion-referenced reports are necessary to understand an individual's score; neither type of report is sufficient.

For example, consider the following reports given to parents of two students. The first parent might be told her student is in the 90th percentile relative to other students in the United States. That statement conveys how well her student scored relative to other students taking the same test. However, that report does not convey what her student knows how to do. It only conveys the student's relative, or normative, position on the test. The second parent might be told that his student understands 42 sight words. This statement conveys information about how well his student has mastered some important prereading skills. This helps tell him what the student has learned, but it does not tell him where the student is relative to others of the same age or level of education. That is, the second parent might not know whether 42 sight words is a good performance or a poor performance relative to other students. Just as the parent who received the first report does not understand what the student can do, the parent who received the second report does not understand where the student is relative to others. Thus, both types of information are necessary to explain a student's score.

Norm-Referenced Scores

LSA provides many norm-referenced scores. The most basic norm-referenced score is a student's rank or standing. The statement "My student finished fifth on a test" is likely to prompt congratulations and a question: "How many others took the test?" However, ranks are cumbersome when large numbers of students are involved. For example, learning your student tied for 14,458th with 229 other students around the country in a year when 36,422 students took the test might tell you exactly how your student compares to the norm group, but it is not easy to understand. Consequently, norm-referenced scores are reported in ways that allow you to understand a student's rank or standing without knowing the number of people who took the test. The most popular types of norm-referenced scores are percentiles, normal curve equivalents, standard scores, and stanines. These are explained in the following sections.

Percentiles

Percentiles are norm-referenced scores between 1 and 99. A percentile represents the proportion, or percentage, of students who scored equal to or worse than the student. A student at the 25th percentile is a student whose score was equal to, or better than, 25% of the students who took the test. Usually, percentiles are reported relative to a national normative group that represents the demographic characteristics of the United States. Some states report scores relative to other students in that state, so that a student's score would be reported as a national percentile and as a state percentile. However, most states are moving away from this practice as they seek to emphasize criterion-referenced scores.

Figure 3.2 shows a typical distribution of scores on a test, with low test scores displayed on the left-hand side and high test scores on the right-hand side. As the line moves from bottom to top, more students are indicated. Therefore, the small space between the bottom line of Figure 3.2 and the curve at the left means few students had very low scores. Likewise, the small space between the curve and the bottom line on the right-hand side means few students had very high scores. The large space between the curve and the line in the middle of the figure means that lots of students have average scores.

Norm-referenced scores are shown in the lines below the curve. Notice that although percentiles are convenient and easily understood, they are not equally spaced. For instance, the difference between students at the 45th percentile and those at the 50th percentile is smaller than the difference between students at the 94th percentile and those at the 99th percentile. In fact, the gap between the 1st and 2nd percentiles is about equal to the difference between the 37th and 50th percentiles. Thus, percentiles give rank order, but they are insensitive to how far apart students are.

Normal Curve Equivalents

A normal curve equivalent (NCE) is a two-digit score also between 1 and 99. However, a normal curve equivalent is an equal-interval scale. It defines how well a student scores relative to the middle of the norm group, and does so in equal units. The middle, or the mean or arithmetic average, of the norm group is set to a score of 50. (Much like the Celsius scale for temperature arbitrarily

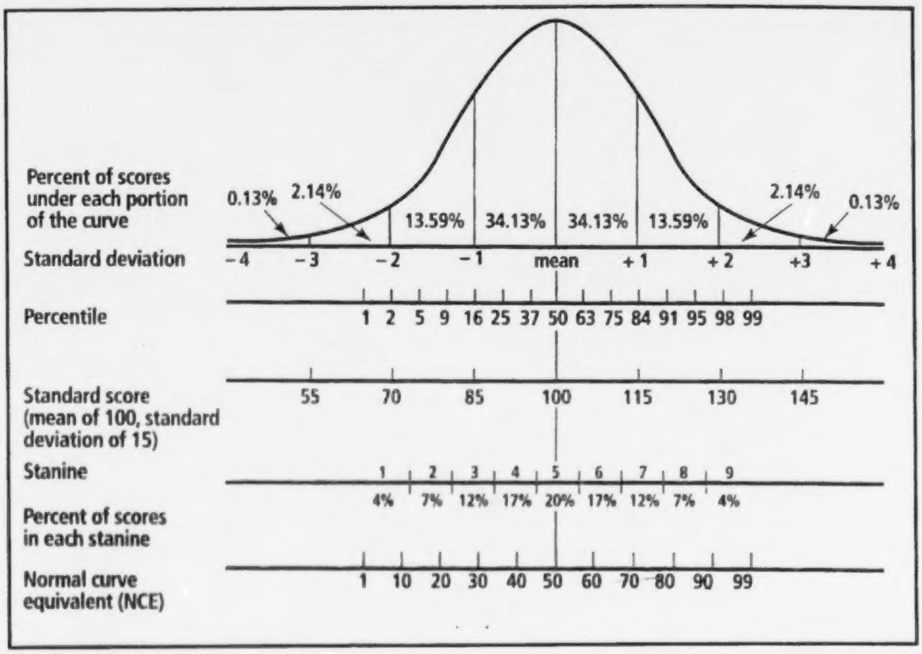

Percent of scores under each portion of the curve
0.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
0.13%
Standard deviation
−4
−3
−2
−1
mean
+1
+2
+3
+4
Percentile
1 2 5 9 16 25 37 50 63 75 84 91 95 98 99
Standard score (mean of 100, standard deviation of 15)
55
70
85
100
115
130
145
Stanine
1 2 3 4 5 6 7 8 9
Percent of scores in each stanine
4% 7% 12% 17% 20% 17% 12% 7% 4%
Normal curve equivalent (NCE)
1 10 20 30 40 50 60 70 80 90 99

FIGURE 3.2
The Normal Curve and Its Relationship to Various Derived Scores

Note. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 27. Copyright February 2000 Wisconsin Department of Public Instruction.

sets 0 to the freezing point of water, and 100 to the boiling point, the NCE scale arbitrarily sets the midpoint of a distribution to 50.) The average spread of individuals about this mean is set to be 21.06. Therefore, a student whose NCE is 30 is about 1 standard deviation below the mean of 50. A student whose NCE is 85 is about 1.75 standard deviations above the mean. Normal curve equivalents are more consistent than percentiles for describing a student's position relative to the norm group, because NCEs are equally spaced. That is, the difference between NCEs of 30 and 35 is the same as the difference between 50 and 55, or 85 and 90 (see for yourself by looking at Figure 3.2). They are better than percentiles, because they reflect position in the norm group using equal units across scores (i.e., NCEs provide rank order and distances between scores). However, normal curve equivalents are not widely understood. Thus, professionals use them to understand students' scores relative to a normative group despite the challenges inherent in understanding them.

Standard Scores

Another way to reflect student scores relative to the norm group is with standard scores. These scores are essentially the same kind of scores as NCEs, but they set the midpoint and standard deviation of the distribution to different values. This is similar to the differences in Celsius and Fahrenheit scales; they each have different values for the freezing point of water (0 versus +32 degrees) and have different spacing between degrees (1 degree on the Celsius scale is nearly 2 degrees on the Fahrenheit scale). Most standard scores fix the middle of the distribution, or mean, to 100, whereas NCEs fix it to 50. Also, the standard deviation of most standard scores is fixed to be 15, versus 21.06 for NCEs. A quick glance at Figure 3.2 shows how the scales compare in describing position on the normal curve. Most group achievement tests use the NCE scale to describe score position, whereas most individually administered tests of intelligence or achievement use the standard score scale to describe score position. The reason for this is strictly habit. Just as you can translate degrees Fahrenheit to degrees Celsius, you can translate standard scores to NCEs, and vice versa, using simple algebra.

Stanines

Stanines are another way to show a score in a form that expresses rank and relative distance between scores. Instead of dividing up the range of scores from 1 to 99 (as NCEs do), or from 55 to 145 (as standard scores do), stanines divide the range of scores into nine equal, or standard, units. (This division is actually how stanines got their name: standard + nine = stanine.) This method simplifies the task of reporting where students are in the distribution, but there is a cost. The intervals between stanines are fairly large, and so stanines are crude, less precise descriptions of student achievement than either NCEs or standard scores. Note that the distance between stanines is constant, except that the lowest (1) and highest (9) stanines are open-ended.

Grade Equivalents

If you use grade equivalents, you may wonder why they are included in this section. Don't grade equivalents describe where a student's score falls in the curriculum? Doesn't a grade equivalent of 3.2 mean a student has mastered the curriculum up to the second month of third grade? Isn't a fourth grader who earns a grade equivalent of 6.8 about 2 to 3 years ahead of curricular expectations? The answer to all of these questions is "No!"

Grade equivalents have little to do with grade-level expectations or with mastery. A grade equivalent is merely the midpoint of a distribution of scores for students in a given grade. To say a score is at the 4.3 grade level is to say the score was equal to the middle score for a group of fourth graders who took the test in the third month of the year (i.e., 4 [grade year] + .3 [month] = 4.3). Grade equivalents are median scores defined so that half of the students in a given grade group will score below the equivalent and half of them will score above the equivalent. In other words, half of all students in the nation are below grade level and, by definition, half are above grade level. No matter how well or poorly our nation's schools educate students, half of all students will be below and half above grade level. Grade equivalent scores are easily misunderstood—that is, most people think they reflect criterion-referenced scores, or mastery of academic subject matter by grade. Because grade equivalent scores are easily misunderstood, we recommend that you do not use them to communicate student progress. The potential for misunderstanding outweighs the potential benefit of understanding. Describe scores relative to a norm using percentiles, NCEs, standard scores, or stanines; avoid using grade equivalents, because they deceive your audience into thinking about curricular comparisons rather than norm comparisons. For this reason, most states no longer report grade equivalents when they describe students' scores. We urge you to just say "No" to grade equivalents! However, we realize that some districts and states require grade equivalents for educational decisionmaking. If you are in one of those districts or states, let your administrator and/or other decisionmakers know of your concerns. We do not advocate insubordination, but we do advocate sharing concerns. Our experience is that many policymakers simply do not understand grade equivalents, and once they do, they embrace a more appropriate score to describe student performance.

Criterion-Referenced Scores

Criterion-referenced scores describe a student's performance relative to a given standard. Popular tests of achievement typically provide four types of criterion-referenced scores: percentages, mastery indexes, scale scores, and proficiency levels. We will describe each of these here, but we encourage you to consult material specific to the test your state or district is using (e.g., Appendixes C, D, and E in this book, guide books for the test) to understand the specific scores your test provides.

Percentages

A percentage is the proportion of items a student answered correctly out of the total number of items in the test. Percentages range from 0% to 100% and are calculated by adding the number of items correct, divided by the total number of items, times 100. Percentages are not percentiles! A student might have 80% correct on a set of items. If the test is difficult, 80% could be a very good score and could result in the student's being in the 99th percentile when compared to others who took the same test. If the test is easy, 80% correct could be a poor score, resulting in the student's being in the 1st percentile when compared to other students who also took the test. Percentages are criterion referenced, because they reflect performance against an absolute (0% to 100%), not normative, standard. However, one caveat is needed: Most tests that report percentage correct do so using estimated scores. That is, they use item response theory and other tools to estimate or predict how a student would have performed on 100 items measuring the same objective, rather than simply reporting the percentage correct (see below).

Mastery Index

These scores (sometimes called *objective performance indicators*, or *OPIs*) estimate the percentage of items a student would get correct in a test in which all items measure the same academic objective or skill. That is, items measuring similar skills within the test are grouped together to measure the academic objectives captured in the test. If there were five items measuring a specific skill (e.g., measurement skills in mathematics), and the student answered four of the items correctly, the student's Mastery Index would be near 80 (i.e., 4/5 × 100). The reason the Mastery Index may not be exactly 80 is that different items are weighted more or less strongly in estimating the Index, based on the items and response characteristics (e.g., a student correctly answering the four easiest of five items gets a lower Mastery Index than the student who answers the four hardest of five items).

Indexes, like percentages, range from 0 to 100. However, they are often grouped into three categories. Each category captures a range of scores. These categories are:

1. Mastery (75–100). Indexes in this range suggest the student has mastered the skill.
2. Partial Mastery (50–74). Indexes in this range suggest the student has partially, but not completely and reliably, mastered the skill.
3. Nonmastery (0–49). Indexes in this range suggest the student has not mastered the skill.

Because Mastery Indexes estimate student mastery of specific curricular skills, they are useful for planning instruction. That is, you could review individual students' scores to identify specific academic strengths and weaknesses if these are reported by your district or state. Likewise, you might review class averages to determine those skills your students have learned and those skills they have not yet mastered, to decide which skills you teach well and which need more instructional attention. (You may want to revisit Chapter 2 and the section on Usability.) It is important to look at two things when considering class-wide results: the mean, or average, Mastery Index and the percentage of students in the class who fall below the mastery level. For example, a class average might be 76 (indicating mastery), yet as many as half the students in the class may fall below mastery level on that skill. We recommend focusing on the proportion of students who have or have not mastered a skill, rather than the average Mastery Index, for making instructional decisions.

Scale Scores

These scores are important to understand and know how to use because they form the basis of all other scores—including state proficiency levels.

To illustrate the concept of scale scores, imagine a curriculum arranged in a line, with one end representing absolutely no knowledge and the other end representing complete mastery of the subject matter domain. Imagine that you put mileposts (like those found on interstate highways) along this line, starting with 0 at the end representing no knowledge and 900 at the end representing mastery. If you had a test in which items were linked to these mile markers, you could use students' responses to test items to estimate how far they had progressed in the curriculum. In fact, this is essentially what most standardized tests of achievement do to yield scale scores. They link specific items to points in the curriculum and place the student along the continuum from 0 to 900. *Note:* Not all tests use 900 as the endpoint. Just as distance can be measured in miles or kilometers, so too can scale scores have different metrics. We will use 900 as an example, because it is the metric used in one of the most popular tests—the TerraNova. However, different tests use different markers. The underlying concept is the same.

Where are students when they enter school on our curricular "highway"? We estimate that most students begin kindergarten at roughly the 400- to 450-mile marker; they have learned nearly half of the subject matter in the curriculum by the time they begin school. Typically, most students have acquired oral language, concepts of numeration, understanding of basic social units, classification skills, and the like before entering kindergarten. Thus, the lowest scale scores typically reported by LSA will be scale scores reflecting about half of their maximum value; the highest scale scores reported on a test will usually be one shy of the maximum (e.g., 899 on a scale of 900). The examinations cannot mark progress for students at or below preschool levels; a student who gets all of the items wrong will still have an estimated scale score in the middle of the scale score range. Therefore, you cannot use LSA to assess students who are working to master early developmental skills, such as toilet skills, feeding, or single-word oral expression. They require an alternate assessment to demonstrate progress in their curriculum.

Scale scores have many advantages over other scores. First, they describe a student's progress in the curriculum regardless of the level of test. For example, a sixth grader whose scale score is 580 would be estimated as having the same level of skills as an eighth grader whose scale score is 580, despite their taking two different levels of the examination. (However, the content of tests varies by level, so they may not be tested on the same set of skills.) Second, scale scores can describe a student's absolute progress in curricula independent of the student's relative standing. For example, a student whose reading scale score from the fourth-grade examination is 510 might be at the 30th percentile relative to other fourth graders. When the same student takes the eighth-grade test, the student's scale score might be 550, but his or her percentile relative to other eighth graders might have dropped to the 10th percentile. The increase in scale scores shows that the student has made progress in the curriculum, but the drop in percentiles shows that the student is not making progress as rapidly as his or her peers. Scale scores provide an absolute, not relative, metric for measuring progress.

The third advantage of scale scores is that they can be used to fix expectations for a given grade level independently of how well other students do on the test. For example, if you were to decide that a scale score of 550 represents what a typical fourth grader should master, you could fix 550 to be a grade-level expectation. It would be statistically possible to have every fourth grader in the nation be at or above this scale score level. Unlike grade equivalents (which rise or fall with the performance of the norm group so that 50% of students are always above or below grade level), scale scores allow educators to fix a standard for grade-level expectations relative to curricular mastery—not the norm group. This is analogous to definitions of physical fitness, in which you might define fitness as the ability to do 10 pull-ups, 50 sit-ups, and 20 push-ups (i.e., set criterion standards), even though the national averages for number of pull-ups (2) , sit-ups (20), and push-ups (7) might fall below your fitness standards. In fact, most states educators use scale scores to define grade-level expectations in the form of proficiency levels.

Proficiency Levels

Proficiency levels set grade-level expectations for curricular mastery. That is, they define certain points in the curriculum (defined by scale score "mile markers") as goals for tests within a subject matter area (e.g., mathematics). How are proficiency levels set? In most cases, proficiency levels are set by test content, not by statistics or scale score properties. The most common procedure for setting standards is "bookmarking."

To set standards using a bookmarking procedure, one begins with a group of people to set standards. Usually, these people are educators, and most are teachers (not politicians). They have content expertise (e.g., language arts, science, mathematics) and grade-level expertise (e.g., they know what fourth, eighth, or tenth graders should be able to know and do). Then, these people get a set of test items in a book, a description of proficiency categories, and a set of bookmarks. The books contain printed items, one per page, arranged from easiest (i.e., the lowest scale score) on the first page to hardest (the highest scale score) on the last page. The proficiency descriptions typically provide general statements of student performance (see Table 3.3 for examples of proficiency descriptions). The set of bookmarks is one less than the number of proficiency descriptions. For example, if there are four proficiency categories, there are three bookmarks; if there are five categories, there are four bookmarks. Each bookmark separates the lower category from the category that is just above it.

Standard setters put a bookmark where they would draw the line between the items that separate performance levels for a given grade. In other words, the standard setter assumes that all items from the first page to the first bookmark reflect items at the lowest proficiency category. They place the second bookmark where they believe the items increase to the next highest category, and so forth. This procedure is usually reiterated several times, with opportunities for standard setters to discuss why they placed their bookmarks where they did. After repeating this process, standard setters eventually come to a consensus regarding the items that define proficiency levels. The scale scores corresponding to the placement of the bookmarks recommended by subject matter/grade level teams define proficiency levels. Most states define proficiency levels for students in a given grade by scale scores on the state LSA.

Note that these scale score levels are set on the basis of item content, or on what students must do to show they have acquired academic skills, not on a statistical basis. It is a rare teacher indeed who knows any test well enough to identify academic content from a scale score alone! Therefore, you might want to better understand the practical meaning of proficiency levels. Here are some activities that can help you become more familiar with proficiency levels and what they mean for your students. These activities take time; you might want to ask your district's inservice/professional development coordinator to set aside time and support the activities.

- Take your state or district test at all levels, or at least at the level nearest your grade. Imagine a student you know fairly well, and who represents about the middle range of skill in your classroom, as you take the test. Answer the items as you think that student might. Be sure also to complete any written essay or extended responses (again, writing as

TABLE 3.3
Sample Proficiency Descriptions for Each Case

Patrick: Florida FCAT Achievement Levels

Level 1	*Level 2*	*Level 3*	*Level 4*	*Level 5*
The student has little success with the challenging content of the Sunshine State Standards.	The student has limited success with the challenging content of the Sunshine State Standards.	The student has partial success with the challenging content of the Sunshine State Standards, but the performance is inconsistent. A Level 3 student answers many of the questions correctly but is generally less successful with the questions that are most challenging.	The student has success with the challenging content of the Sunshine State Standards. A Level 4 student answers most of the questions correctly but may have only some success with questions that reflect the most challenging content.	The student has success with the most challenging content of the Sunshine State Standards. A Level 5 student answers most of the questions correctly, including the most challenging questions.

Tia: Wisconsin's Proficiency Categories

Minimal Performance	*Basic*	*Proficient*	*Advanced*
Limited achievement in the academic knowledge and skills tested.	Somewhat competent in the academic knowledge and skills tested.	Competent in the important academic knowledge and skills tested.	Distinguished achievement. In-depth understanding of academic knowledge and skills tested.

Chris: Idaho's Performance Levels*

Below Basic	*Basic***	*Proficient***	*Advanced***
Students . . . do not meet the grade level standard for basic achievement.	Students . . . understand the overall literal meaning of the text that they read.	Students . . . identify ideas and information suggested by, but not explicitly stated in, the text that they read.	Students . . . generalize about ideas and information in the text that they read and evaluate the texts critically.

* Idaho uses the levels provided by Riverside Publishing for the Iowa Tests of Basic Skills.

** The definition of Basic, Proficient, and Advanced categories are subject and grade specific. Those in the table are for fourth-grade reading.

your student might). You can usually get copies of past examinations given to students from your district assessment coordinator. You may not make copies of these assessments, and you should clearly state that your intent is to learn about the test generally and proficiencies in particular. You should not try to copy or memorize items to teach to students, because it is unethical (and usually a waste of time, since test items always change from year to year).

- Score your examination. Use the scoring guide for the level(s) of test you took, and score your responses. Some responses are scored easily, whereas others require judgment. For example, many tests require you to determine the differences among one-, two-, and three-point responses on a short written answer, and in some cases you will score the same response twice (e.g., once for grammar/style and once for content/meaning). You may have to score your essay using a holistic rating rubric or using anchor papers. Then ask a colleague to score the essay. Do not tell the colleague how you scored it. Compare your essay scoring to your colleague's scoring of the same one. If the scores are identical, that is the final score for your essay; if the scores are within one point of each other, simply add the two scores and divide by two. If the scores are more than one point apart, get another colleague to score the essay. Add the two closest scores and divide by two to get the final essay score. Your district assessment coordinator can usually supply you with scoring guides to help you with this activity. This makes an excellent inservice activity for groups of teachers.

These exercises will help you better understand the content of the LSA used in your district and state and how that content is linked to district and state standards. Knowledge of test content is a necessary, but not sufficient, condition for making informed and effective judgments about what, how, and when to teach material. Also, knowledge of the examinations is essential for deciding whether and how students with disabilities should participate in LSA.

Applying Your Knowledge of LSA: Looking at Tia

We now provide an in-depth look at the kinds of reports that states and districts produce from the results of LSA. We do not have enough space to present and discuss all of the ways in which districts and states present results; consequently, we provide an in-depth look at one state as an example of the kinds of reports produced from LSA. We selected Tia's state—Wisconsin—to elaborate how to interpret LSA results.

Because Tia is in eighth grade, she should participate in the Wisconsin Knowledge and Concepts Examinations (WKCE), which are based on the TerraNova. Wisconsin administers the WKCE/TerraNova at fourth, eighth, and tenth grades to cover reading, language arts, mathematics, social studies, and science. Most states also test at elementary, middle, and high school levels to remain eligible for ESEA Title I funds, although some states test fewer subjects (e.g., language arts and mathematics) or stagger subjects across different grades (e.g., third-grade mathematics and science; fourth-grade reading and language arts).

Let's use your knowledge of LSA scores to interpret the results of the WKCE. The WKCE reports results in many ways and for many targets (e.g., individual students, classrooms, districts). Our in-depth look at Tia will guide you in interpreting the following reports:

- Individual Profile Report.
- Group Proficiency Level Report.
- Evaluation Summary Report.
- School Record Sheet.
- Writing Frequency Distribution.
- Objectives Performance Summary.

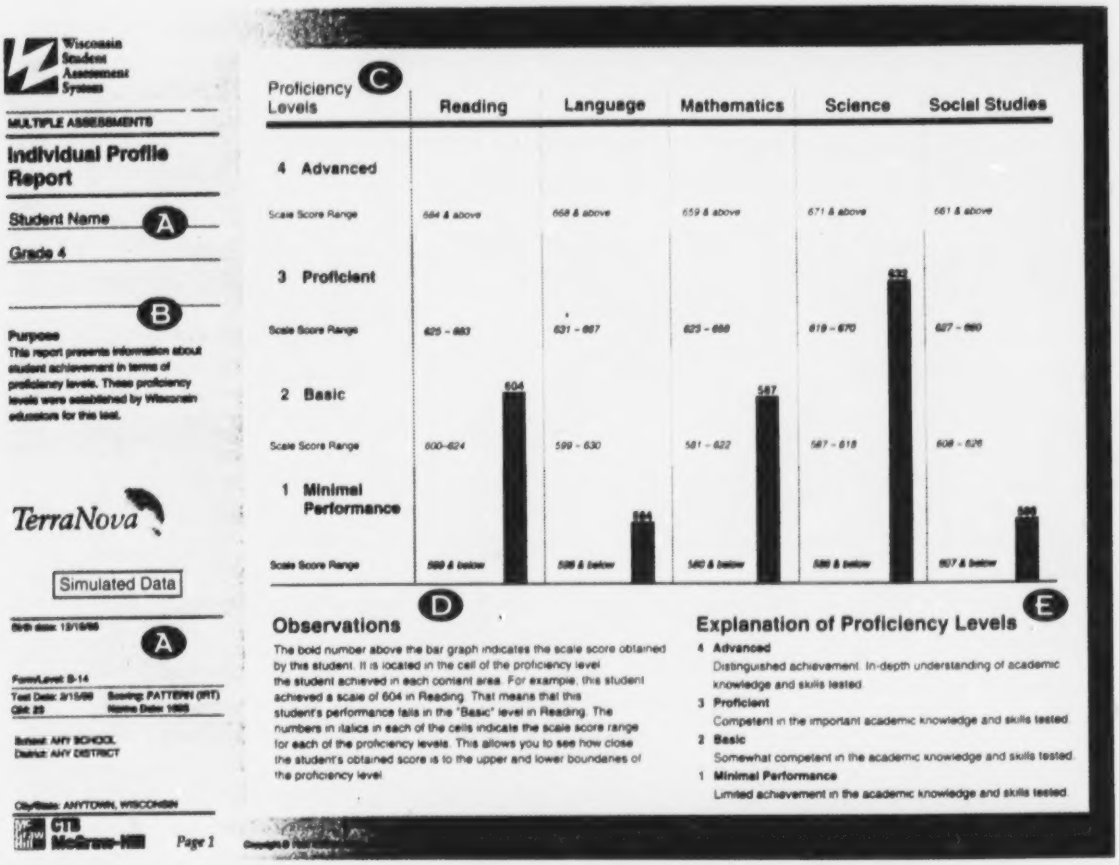
Wisconsin generates one other type of report for each district, the Item Analysis Summary. Because it is used primarily by district assessment specialists, and not by teachers, we will not describe it in this chapter.

Individual Profile Report

An example of a fourth-grade student's Individual Profile Report appears in Figure 3.3. You will note that the report is two pages long. On the first page, the student's proficiency level in five subject matter areas (i.e., Reading, Language, Mathematics, Science, Social Studies) is presented in graph form. For example, this student's achievement in Reading was at the Basic proficiency level, but the student's achievement in Language was at the Minimal Performance level.

The top section of the report's second page describes the student's results using stanines, scale scores, and national percentiles. Look at Figure 3.3 to see how the student's stanines and percentiles compare to those of others. In all areas, the student is above the average for students taking the test. Compare the student's scale scores to the scale score proficiency ranges on page 1 of the report. You can see that the student's scale scores meet or exceed the lowest boundary of the Proficient range in Science (i.e., the scale score of 632 is between 619 and 670). The student's scale scores in Reading and Mathematics are in the Basic proficiency level, whereas the student's scores in Language and Social Studies fall below Basic (i.e., reflect Minimal Performance). Finally, note that the last column of the section reports a National Percentile Range for each of the student's scores. This range uses the estimated likelihood of error in the score (remember, no test is perfect!) to predict where the student's performance actually falls. For example, your best estimate for the student's percentile rank in Language is 53, but you know there is some error in the test, so you would be pretty confident that the student's "true" percentile would fall between the 43rd and 60th percentiles.

The bottom section of page 2 of the report tells the type of prompt (Informative, Narrative, Descriptive, or Persuasive) given the student. The holistic score of 4.5 tells you one rater scored the essay a 4 and the other scored


Wisconsin Student Assessment System
A
B
C
D
E
TerraNova
CTB McGraw-Hill
Page 1
Proficiency Levels
Reading
Language
Mathematics
Science
Social Studies
4 Advanced
Scale Score Range
684 & above
668 & above
659 & above
671 & above
661 & above
3 Proficient
Scale Score Range
625 - 683
631 - 667
623 - 658
619 - 670
627 - 660
2 Basic
Scale Score Range
600-624
599 - 630
581 - 622
587 - 618
608 - 626
1 Minimal Performance
Scale Score Range
599 & below
598 & below
580 & below
586 & below
607 & below
604
584
587
632

MULTIPLE ASSESSMENTS

Individual Profile Report

Student Name

Grade 4

Purpose
This report presents information about student achievement in terms of proficiency levels. These proficiency levels were established by Wisconsin educators for this test.

Simulated Data

Form/Level: B-14
Scoring: PATTERN (IRT)

School: ANY SCHOOL
District: ANY DISTRICT

City/State: ANYTOWN, WISCONSIN

Observations

The bold number above the bar graph indicates the scale score obtained by this student. It is located in the cell of the proficiency level the student achieved in each content area. For example, this student achieved a scale of 604 in Reading. That means that this student's performance falls in the "Basic" level in Reading. The numbers in italics in each of the cells indicate the scale score range for each of the proficiency levels. This allows you to see how close the student's obtained score is to the upper and lower boundaries of the proficiency level.

Explanation of Proficiency Levels

4 **Advanced**
Distinguished achievement. In-depth understanding of academic knowledge and skills tested.
3 **Proficient**
Competent in the important academic knowledge and skills tested.
2 **Basic**
Somewhat competent in the academic knowledge and skills tested.
1 **Minimal Performance**
Limited achievement in the academic knowledge and skills tested.

FIGURE 3.3
Individual Profile Report

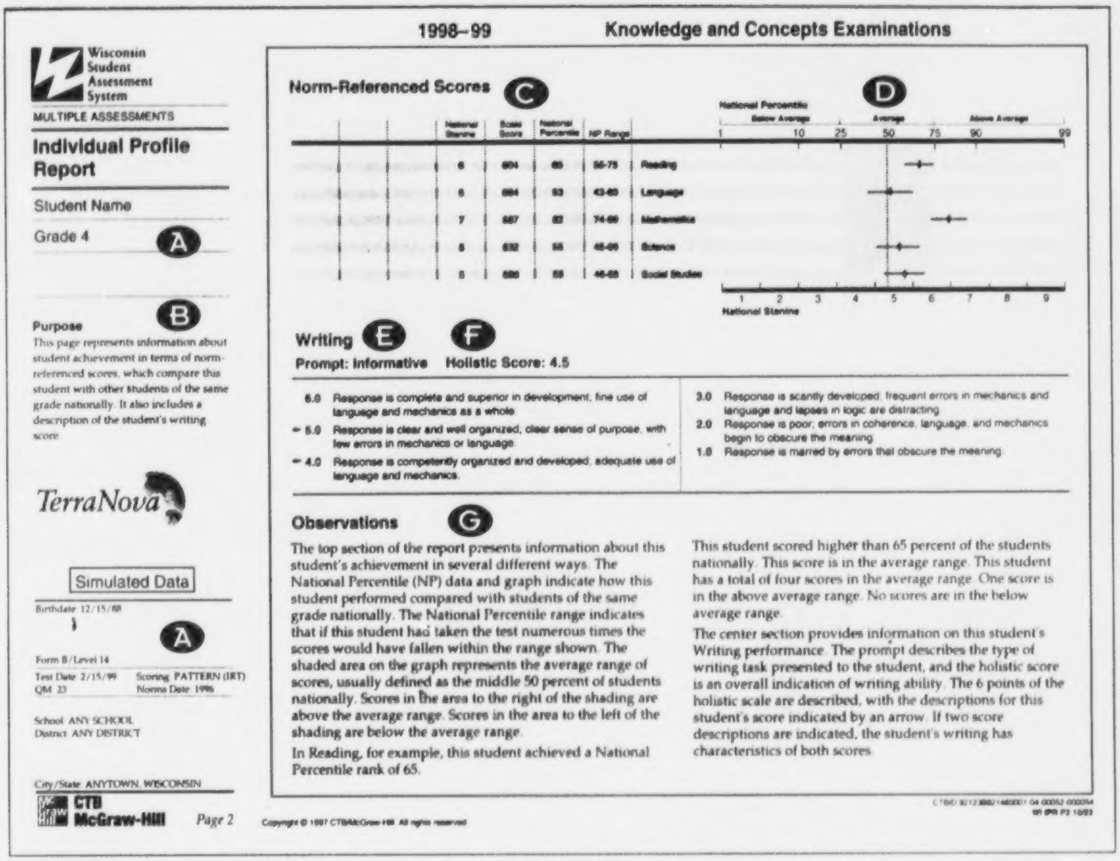
From *Wisconsin Student Assessment System* (based on the Terra/Nova tests), 1997, Monterey, California: CTB/McGraw-Hill. Copyright 1997 by CTB/McGraw-Hill. Reproduced with permission of CTB/McGraw-Hill.

it a 5, yielding a final score of 4.5 (i.e., (4 + 5)/2 = 4.5). The descriptions below the score describe the essay quality.

Group Proficiency Level Report

Figure 3.4 presents a Group Proficiency Level Report for a fourth-grade class of 30 students. This report describes the proportion of students in each proficiency category for the class, school, district, and state (rows) by subject matter area (columns). Looking at the top row of the second column (Reading), you can see that 27 students (of 30) took the Reading test. Within the Reading domain, 16, or 53%, of the students' scores fell in the Minimal Performance range. This compares to 49% of scores for fourth graders at that school, 45% of fourth graders in the district, and 44% of fourth graders across the state. None of the students in this class scored at the Proficient or Advanced level on the Reading test. In contrast, 21 of 30 (70%) scored at the Proficient level in Science.

BEST COPY AVAILABLE


National Percentile
Below Average
Average
Above Average
1
10
25
50
75
90
99
1
2
3
4
5
6
7
8
9
National Stanine
A
B
C
D
E
F
G
A

Wisconsin Student Assessment System

MULTIPLE ASSESSMENTS

Individual Profile Report

Student Name

Grade 4

Purpose

This page represents information about student achievement in terms of norm-referenced scores, which compare this student with other students of the same grade nationally. It also includes a description of the student's writing score.

TerraNova

Simulated Data

Birthdate 12/15/88

Form B/Level 14

Test Date 2/15/99 Scoring PATTERN (IRT)
QM 23 Norms Date 1996

School ANY SCHOOL
District ANY DISTRICT

City/State ANYTOWN, WISCONSIN

CTB McGraw-Hill Page 2

1998–99 Knowledge and Concepts Examinations

Norm-Referenced Scores

National Stanine	Scale Score	National Percentile	NP Range	
6	604	65	56-75	Reading
5	594	53	43-60	Language
7	637	82	74-89	Mathematics
5	632	56	46-65	Science
5	590	58	46-68	Social Studies

Writing

Prompt: Informative **Holistic Score: 4.5**

6.0 Response is complete and superior in development, fine use of language and mechanics as a whole.
→ 5.0 Response is clear and well organized, clear sense of purpose, with few errors in mechanics or language.
→ 4.0 Response is competently organized and developed; adequate use of language and mechanics.
3.0 Response is scantly developed; frequent errors in mechanics and language and lapses in logic are distracting.
2.0 Response is poor; errors in coherence, language, and mechanics begin to obscure the meaning.
1.0 Response is marred by errors that obscure the meaning.

Observations

The top section of the report presents information about this student's achievement in several different ways. The National Percentile (NP) data and graph indicate how this student performed compared with students of the same grade nationally. The National Percentile range indicates that if this student had taken the test numerous times the scores would have fallen within the range shown. The shaded area on the graph represents the average range of scores, usually defined as the middle 50 percent of students nationally. Scores in the area to the right of the shading are above the average range. Scores in the area to the left of the shading are below the average range.

In Reading, for example, this student achieved a National Percentile rank of 65.

This student scored higher than 65 percent of the students nationally. This score is in the average range. This student has a total of four scores in the average range. One score is in the above average range. No scores are in the below average range.

The center section provides information on this student's Writing performance. The prompt describes the type of writing task presented to the student, and the holistic score is an overall indication of writing ability. The 6 points of the holistic scale are described, with the descriptions for this student's score indicated by an arrow. If two score descriptions are indicated, the student's writing has characteristics of both scores.

Copyright © 1997 CTB/McGraw-Hill. All rights reserved.

FIGURE 3.3
Individual Profile Report *Continued*

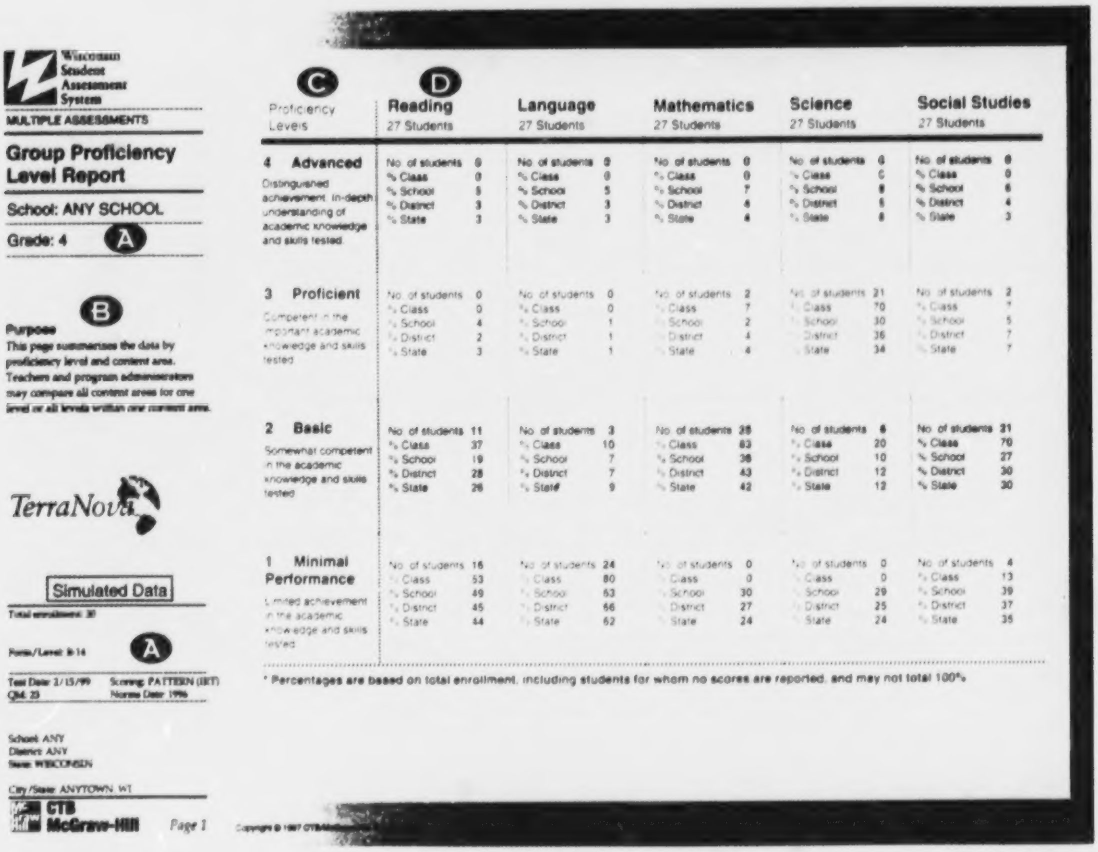
From *Wisconsin Student Assessment System* (based on the Terra/Nova tests), 1997, Monterey, California: CTB/McGraw-Hill. Copyright 1997 by CTB/McGraw-Hill. Reproduced with permission of CTB/McGraw-Hill.

Note that the percentage proficient can be based on the proportion of students who are enrolled in a given grade or the number of students who took the test. These are usually different numbers, and consequently, yield different percentages in each proficiency category. Read "Playing the Percentages" (Figure 3.5) to learn more about how students are counted in proficiency categories and why it can make a big difference in the results.

Evaluation Summary Report

The Evaluation Summary Report describes the achievement scores for a school or district at fourth, eighth, or tenth grade. Figure 3.6 contains an example of an Evaluation Summary Report for "Any School's" class of 89 (see lower left-hand side of the report) eighth-grade (see letter A) students. The top row of results tells the number of students whose scores are included in the summary

BEST COPY AVAILABLE



Wisconsin Student Assessment System

MULTIPLE ASSESSMENTS

Group Proficiency Level Report

School: ANY SCHOOL

Grade: 4

Purpose
This page summarizes the data by proficiency level and content area. Teachers and program administrators may compare all content areas for one level or all levels within one content area.

TerraNova

Simulated Data

Total enrollment: 30

Form/Level: B-14

Test Date: 2/15/99
QM: 23
Scoring: PATTERN (IRT)
Norms Date: 1996

School: ANY
District: ANY
State: WISCONSIN

City/State: ANYTOWN, WI

CTB McGraw-Hill

Page 1

Copyright © 1997 CTB

Proficiency Levels		Reading 27 Students	Language 27 Students	Mathematics 27 Students	Science 27 Students	Social Studies 27 Students
4 Advanced Distinguished achievement. In-depth understanding of academic knowledge and skills tested.	No. of students	0	0	0	0	0
	% Class	0	0	0	0	0
	% School	5	5	7	8	6
	% District	3	3	4	5	4
	% State	3	3	4	8	3
3 Proficient Competent in the important academic knowledge and skills tested	No. of students	0	0	2	21	2
	% Class	0	0	7	70	7
	% School	4	1	2	30	5
	% District	2	1	4	36	7
	% State	3	1	4	34	7
2 Basic Somewhat competent in the academic knowledge and skills tested	No. of students	11	3	25	6	21
	% Class	37	10	83	20	70
	% School	19	7	38	10	27
	% District	28	7	43	12	30
	% State	26	9	42	12	30
1 Minimal Performance Limited achievement in the academic knowledge and skills tested	No. of students	16	24	0	0	4
	% Class	53	80	0	0	13
	% School	49	63	30	29	39
	% District	45	66	27	25	37
	% State	44	62	24	24	35

* Percentages are based on total enrollment, including students for whom no scores are reported, and may not total 100%

FIGURE 3.4
Proficiency Summary by Student Group

From *Wisconsin Student Assessment System* (based on the Terra/Nova tests), 1997, Monterey, California: CTB/McGraw-Hill. Copyright 1997 by CTB/McGraw-Hill. Reproduced with permission of CTB/McGraw-Hill.

report. Note that the number varies by subject matter, with only 86 students completing the Reading section and 89 completing the Mathematics section.

The second major row (letter C) of results lists the arithmetic average, or mean, for many scores and the average spread of scores about the mean (the standard deviation). Each line in this row is described in the following list:

- The top line provides the mean, or average, NCE for the five subject matter areas. Examples: the mean Reading NCE for this class is 48.0; the mean Science NCE is 51.4. Remember: 50 is the national mean, so all of these scores are close to the national average.
- The second line reports the average spread (i.e., standard deviation) around the mean. Examples: the average spread of Reading NCEs around the mean is 13.9; the average spread of NCEs in Mathematics is bigger (19.3). Remember: a representative normal sample would be about 21; standard deviations of less than 16 imply that students are more alike than would be expected, and numbers greater than 26 suggest that students are more diverse than expected.

BEST COPY AVAILABLE

Where do the percentages in a Group Proficiency Level Report come from? The answer is not obvious. To answer the question, look closely at Figure 3.4

First, note that 3, or 10%, of fourth graders in this class did not take the WKCE. Reasons for not taking the test might include limited English proficiency, poor attendance, or exclusion due to disabilities. That means that 27, or 90%, of the fourth graders in this class took the exams, and 3, or 10%, did not.

Second, the percentage of students in each category is based on the total number of students enrolled in the class. In this example, the enrollment number was 30 students. To calculate the proportion of students in each proficiency level, the report takes the number who scored at that level (e.g., 21) scored at the Proficient level in Science) and divides by the total enrolled (30) to get the proportion of students in the class at the Proficient level (70% = 21/30).

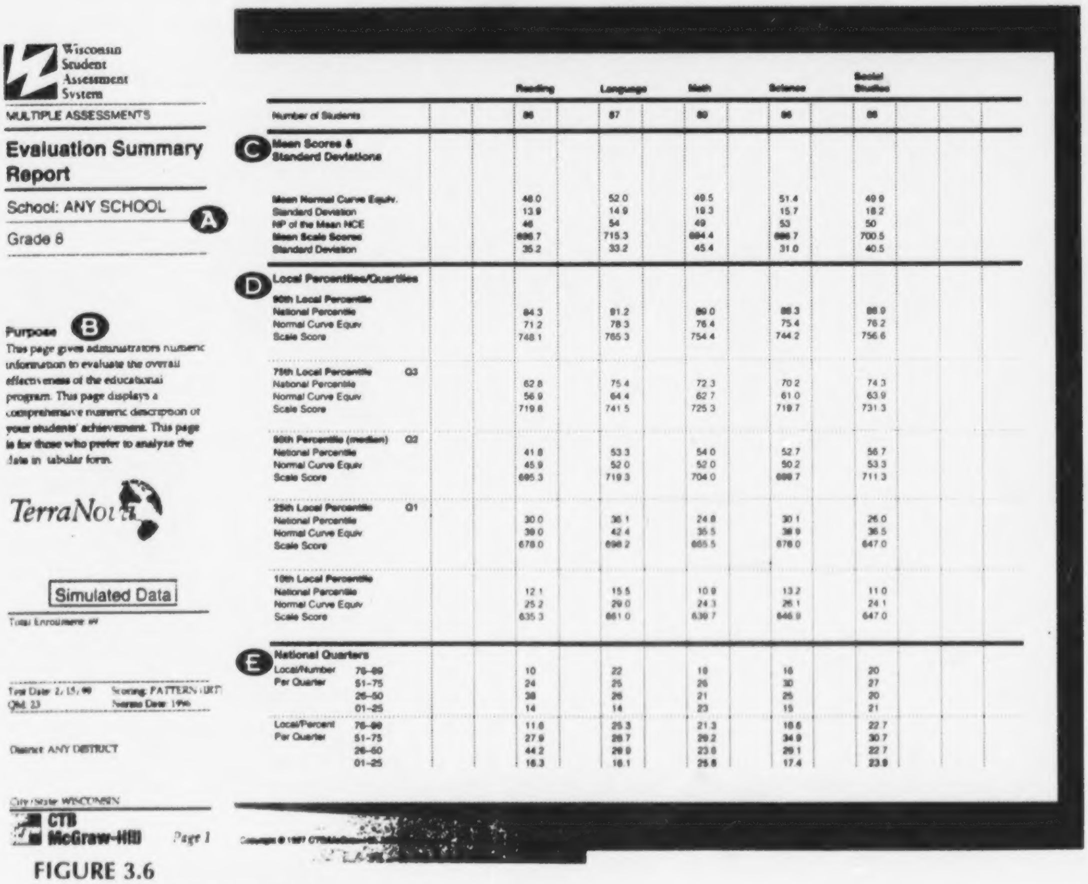
You might argue that the results underestimate the percentage of children in this class who are in a given proficiency level. For example, you could say that 100% of the students who took the Science test scored at or above the Basic proficiency level (i.e., 6 + 21 = 27, or 100% of testtakers). You could say the same for Mathematics. However, the report shows that only 90% of the students scored at or above the Basic level in Science and Mathematics. Why isn't it 100%?

The answer lies in "playing percentages." By reporting results as a proportion of students who are enrolled, rather than the proportion of students who took the test, the state is eliminating incentives for excluding students from the state test. If a state reported outcomes in terms of the proportion who took the test (rather than total enrollment), it might encourage districts to exclude the lowest-scoring students from the WKCE. For example, if you excluded the 6 students who scored at the Basic level in Science, plus the 3 students who did not take the exam, 21 students would score in the Proficient level. That would mean 21/21 students, or 100% of those taking the test, would be Proficient! However, only 21 (i.e., 70% of the class) actually earned scores at the Proficient or Advanced levels. So, the percentage of students in each proficiency category is determined by the number of students who earn scores in that category, divided by the number of students enrolled in the grade—not by the number of students who took the test. The state reported the percentage at each proficiency level based on the total in the class, rather than the total who took the test, so that districts would not be inadvertently encouraged to exclude students who might score lower than others. Schools have nothing to lose—and perhaps something to gain—by including students in the state test.

Check whether your state uses the total who took the test or the total enrolled to calculate proportions of students falling at given levels or categories. It makes a difference in how motivated schools may be to include students with disabilities in state tests.

FIGURE 3.5
Playing the Percentages

- The third line reports the national percentile (NP) of the NCE mean. Examples: the mean Science NCE of 51.4 is equal to an NP of 53; the Social Studies NCE mean of 49.9 is equal to an NP of 50. Remember: the average NP is 50 (i.e., an NP of 50 divides the national sample in half, with half scoring lower and half scoring higher).
- The fourth line reports the mean scale score for the group. Examples: the mean scale score for Science (696.7) is lower than the mean scale score for Social Studies (700.5). Remember: scale scores are like yardsticks, so it is possible to compare scores across academic domains and different levels of the test. The fact that the Social Studies NP is lower



Wisconsin Student Assessment System

MULTIPLE ASSESSMENTS

Evaluation Summary Report

School: ANY SCHOOL Ⓐ

Grade 8

Purpose Ⓑ

This page gives administrators numeric information to evaluate the overall effectiveness of the educational program. This page displays a comprehensive numeric description of your students' achievement. This page is for those who prefer to analyze the data in tabular form.

TerraNova

Simulated Data

Total Enrollment: ##

Test Date: 2/15/99 Scoring: PATTERN (IRT)
QM: 23 Norms Date: 1996

District: ANY DISTRICT

City/State: WISCONSIN

CTB McGraw-Hill Page 1

		Reading	Language	Math	Science	Social Studies
Number of Students		86	87	88	86	88
Ⓒ **Mean Scores & Standard Deviations**						
Mean Normal Curve Equiv.		48.0	52.0	49.5	51.4	49.9
Standard Deviation		13.9	14.9	19.3	15.7	18.2
NP of the Mean NCE		46	54	49	53	50
Mean Scale Scores		686.7	715.3	694.4	686.7	700.5
Standard Deviation		35.2	33.2	45.4	31.0	40.5
Ⓓ **Local Percentiles/Quartiles**						
90th Local Percentile						
National Percentile		84.3	91.2	89.0	88.3	88.9
Normal Curve Equiv		71.2	78.3	76.4	75.4	76.2
Scale Score		748.1	765.3	754.4	744.2	756.6
75th Local Percentile	Q3					
National Percentile		62.8	75.4	72.3	70.2	74.3
Normal Curve Equiv		56.9	64.4	62.7	61.0	63.9
Scale Score		719.8	741.5	725.3	719.7	731.3
50th Percentile (median)	Q2					
National Percentile		41.8	53.3	54.0	52.7	56.7
Normal Curve Equiv		45.9	52.0	52.0	50.2	53.3
Scale Score		695.3	719.3	704.0	698.7	711.3
25th Local Percentile	Q1					
National Percentile		30.0	36.1	24.8	30.1	26.0
Normal Curve Equiv		39.0	42.4	35.5	38.9	36.5
Scale Score		678.0	698.2	665.5	678.0	647.0
10th Local Percentile						
National Percentile		12.1	15.5	10.9	13.2	11.0
Normal Curve Equiv		25.2	29.0	24.3	26.1	24.1
Scale Score		635.3	661.0	630.7	646.9	647.0
Ⓔ **National Quarters**						
Local/Number Per Quarter	76–99	10	22	18	16	20
	51–75	24	25	26	30	27
	26–50	38	26	21	25	20
	01–25	14	14	23	15	21
Local/Percent Per Quarter	76–99	11.6	25.3	21.3	18.6	22.7
	51–75	27.9	28.7	29.2	34.9	30.7
	26–50	44.2	29.9	23.8	29.1	22.7
	01–25	16.3	16.1	25.8	17.4	23.9

Copyright © 1997 CTB/McGraw-

FIGURE 3.6
Evaluation Summary Report

From *Wisconsin Student Assessment System* (based on the Terra/Nova tests), 1997, Monterey, California: CTB/McGraw-Hill. Copyright 1997 by CTB/McGraw-Hill. Reproduced with permission of CTB/McGraw-Hill.

than the Science NP (even though the scale score is higher) means the national sample finds social studies easier than science.

- The fifth line reports the average spread (i.e., standard deviation) of scores around the scale score mean. Examples: the spread of Language scale scores (33.2) is smaller than the spread of Mathematics scale scores (45.4).

The next major section or row of the Evaluation Summary (letter D) divides the group of scores into different sections. The sections are defined by the score that separates the top 10% from the rest of the class (i.e., the 90th Local Percentile, or LP); the score separating the top 25% (75th LP); the median for the class (50th LP), the score separating the bottom 25% (25th LP), and the bottom 10% (10th LP). This information tells you how scores are spread out—or bunched up—within a class. Within each of these sections, there are three lines reporting results:

- The first line reports the National Percentile (NP) of the LP. Examples: the score defining the top 10% of the class (90th LP) is equal to an NP

BEST COPY AVAILABLE

of 91.2 in Language and 84.3 in Reading. The median class score (50th LP) in Science has an NP of 52.7, and the median Reading score is 41.8. Remember: in a class that exactly reflects the national average, the NP of the 50th LP (median) would be 50 (i.e., the score defining the top 50% would be the same for the class and the national average); in classes that score higher than the average, the NP will be over 50, and in classes below the national average, the median NP will be less than 50.

- The second line reports the NCE of the LP. Examples: the NCE of the bottom 10 percent of the class in Mathematics is 24.3; the Science NCE for the top quarter (75th LP) is 61.0.
- The third line reports the scale score of the LP. Example: the 25th LP (bottom quarter of the class) is defined by a Language scale score of 698.2.

The bottom row or section of the report (letter E) tells you how many students had scores within the top, second, third, and bottom quarters relative to national averages.

- The first row of four lines tells the number of students in the class within each national quartile. Examples: 14 students scored in the bottom quartile on the Reading test, whereas 23 students scored in the bottom quartile of the Math test; on the Language test, 22 students scored in the top quartile and 25 scored in the second quartile. Remember: the number of students in any quartile is determined by how well they do on the test and by the number who took the test.
- The second row of four lines tells the percentage or proportion of students in the class within each national quartile. Examples: 11.6% of the class placed in the top quartile in Reading; 25.3% of the class placed in the top quartile in Language. Remember: the proportion expected in each national quartile is 25%. If the proportion in the top two quartiles is greater than 50%, the class is above the national average; if the numbers add to less than 50%, the class is below the national average. This is true no matter how many students take the test (25% is always expected in each quartile).

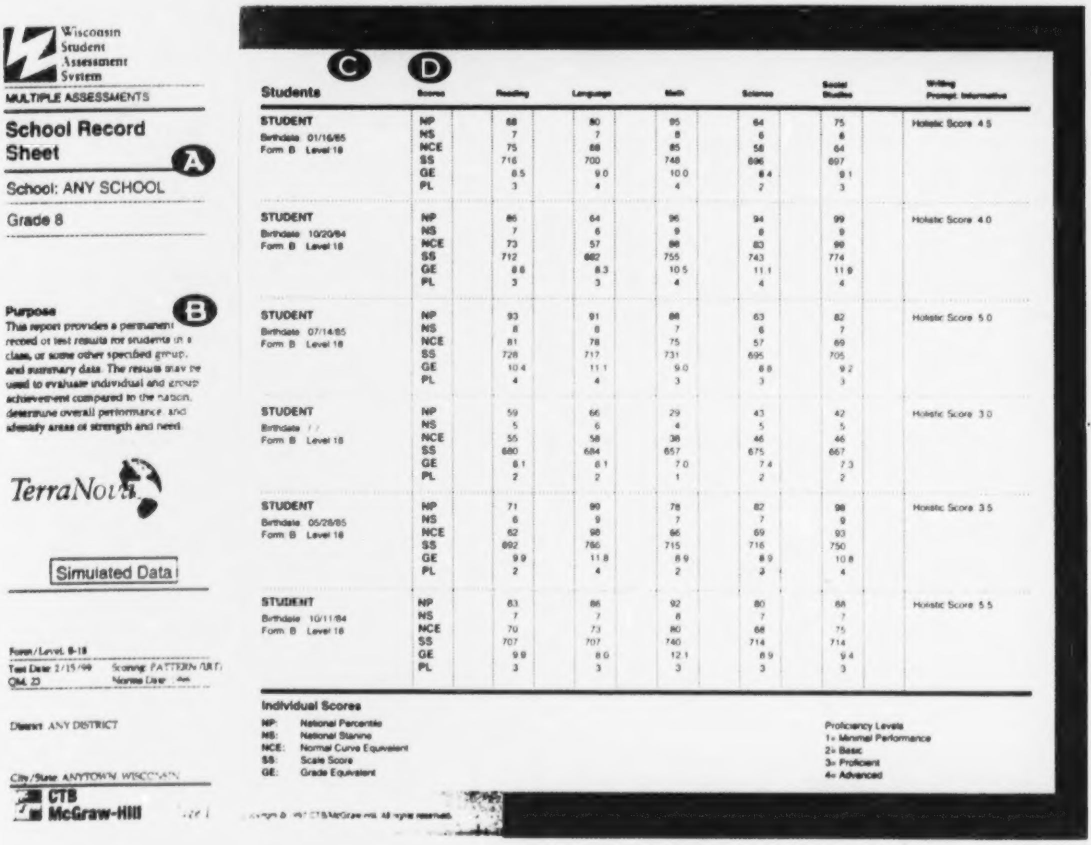
School Record Sheet

This document lists each student's scores in each academic domain. Figure 3.7 presents the first page (page 1) of scores from a group of eighth graders, and the last page (page 2) of scores from a group of fourth graders. Each row represents a different student (on page 1) and a final proficiency summary for all students (page 2).

- The first column of the report (letter C) identifies the students by name (omitted on this report), birth date, and the form of the test the students took (Form B, Level 18).
- The second column (letter D) lists the scores reported for each student. They are:

  NP (National Percentile; range 1–99).

  NS (National Stanine, range 1–9).



Wisconsin Student Assessment System

MULTIPLE ASSESSMENTS

School Record Sheet

School: ANY SCHOOL

Grade 8

Purpose

This report provides a permanent record of test results for students in a class, or some other specified group, and summary data. The results may be used to evaluate individual and group achievement compared to the nation, determine overall performance, and identify areas of strength and need.

TerraNova

Simulated Data

Form/Level: B-18

Test Date: 2/15/99 Scoring: PATTERN (IRT)

QM: 23 Norms Date: 1996

District: ANY DISTRICT

City/State: ANYTOWN, WISCONSIN

CTB McGraw-Hill

Students	Scores	Reading	Language	Math	Science	Social Studies	Writing Prompt: Informative
STUDENT Birthdate 01/16/85 Form B Level 18	NP	88	80	95	64	75	Holistic Score 4.5
	NS	7	7	8	6	6	
	NCE	75	68	85	58	64	
	SS	716	700	748	696	697	
	GE	8.5	9.0	10.0	8.4	9.1	
	PL	3	4	4	2	3	
STUDENT Birthdate 10/20/84 Form B Level 18	NP	86	64	96	94	99	Holistic Score 4.0
	NS	7	6	9	8	9	
	NCE	73	57	88	83	99	
	SS	712	682	755	743	774	
	GE	8.8	8.3	10.5	11.1	11.9	
	PL	3	3	4	4	4	
STUDENT Birthdate 07/14/85 Form B Level 18	NP	93	91	88	63	82	Holistic Score 5.0
	NS	8	8	7	6	7	
	NCE	81	78	75	57	69	
	SS	728	717	731	695	705	
	GE	10.4	11.1	9.0	6.8	9.2	
	PL	4	4	3	3	3	
STUDENT Birthdate / / Form B Level 18	NP	59	66	29	43	42	Holistic Score 3.0
	NS	5	6	4	5	5	
	NCE	55	58	38	46	46	
	SS	680	684	657	675	667	
	GE	8.1	8.1	7.0	7.4	7.3	
	PL	2	2	1	2	2	
STUDENT Birthdate 06/28/85 Form B Level 18	NP	71	90	78	82	98	Holistic Score 3.5
	NS	6	9	7	7	9	
	NCE	62	98	66	69	93	
	SS	692	786	715	716	750	
	GE	9.9	11.8	8.9	8.9	10.8	
	PL	2	4	2	3	4	
STUDENT Birthdate 10/11/84 Form B Level 18	NP	83	86	92	80	88	Holistic Score 5.5
	NS	7	7	8	7	7	
	NCE	70	73	80	68	75	
	SS	707	707	740	714	714	
	GE	9.9	8.0	12.1	8.9	9.4	
	PL	3	3	3	3	3	

Individual Scores

NP:	National Percentile
NS:	National Stanine
NCE:	Normal Curve Equivalent
SS:	Scale Score
GE:	Grade Equivalent

Proficiency Levels
1= Minimal Performance
2= Basic
3= Proficient
4= Advanced

Copyright © 1997 CTB/McGraw-Hill. All rights reserved.

FIGURE 3.7
School Record Sheet

From *Wisconsin Student Assessment System* (based on the Terra/Nova tests), 1997, Monterey, California: CTB/McGraw-Hill. Copyright 1997 by CTB/McGraw-Hill. Reproduced with permission of CTB/McGraw-Hill.

NCE (Normal Curve Equivalent; range 1–99).

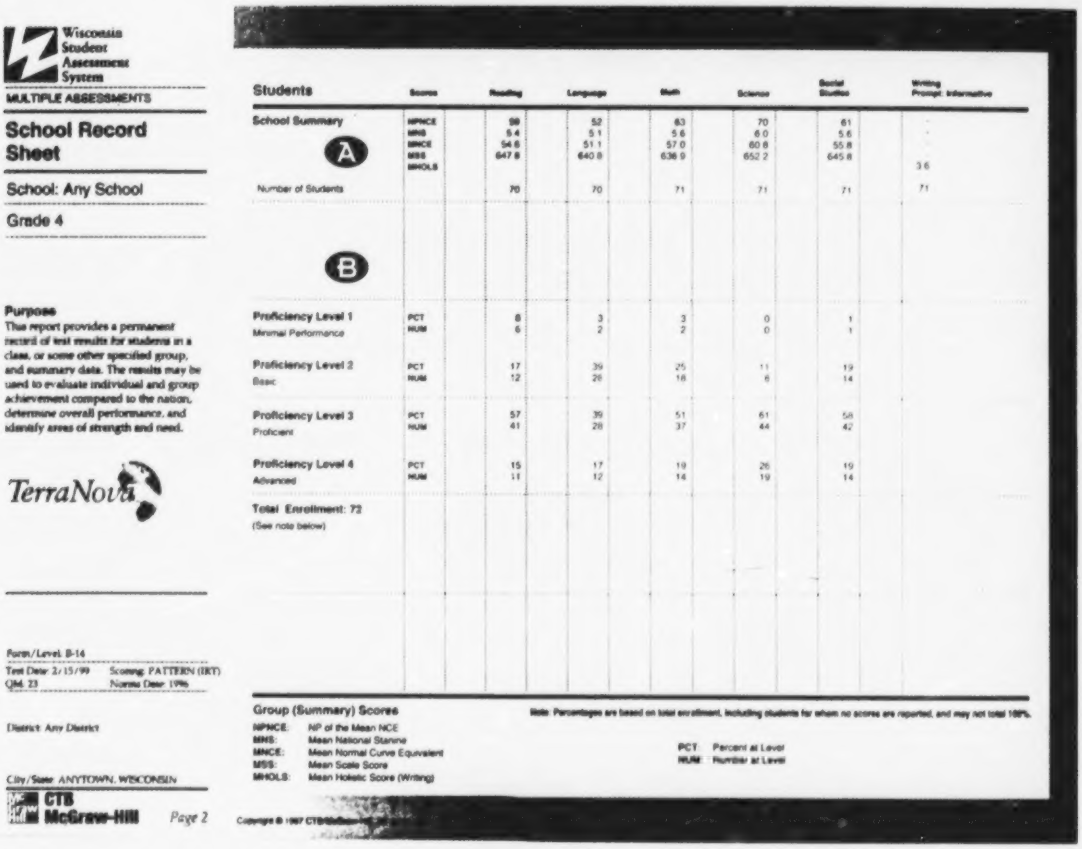
SS (Scale Score; range 450–899).

GE (Grade Equivalent; range pre-K–12.9+).

PL (Proficiency Level; 1 = Minimal Performance, 2 = Basic, 3 = Proficient, 4 = Advanced).

- The next column presents each student's Reading score in six different ways (NP, NS, NCE, SS, GE, PL).
- The next four columns present each student's scores in Language, Math, Science, and Social Studies in six different ways (NP, NS, NCE, SS, GE, PL).
- The last column presents each student's holistic writing score (all students responded to the Informative Writing Prompt), which ranges from 1 to 6 (see Figure 3.3, page 2 of the Individual Score Report, for descriptions of each score).

BEST COPY AVAILABLE



Wisconsin Student Assessment System

MULTIPLE ASSESSMENTS

School Record Sheet

School: Any School

Grade 4

Purpose
This report provides a permanent record of test results for students in a class, or some other specified group, and summary data. The results may be used to evaluate individual and group achievement compared to the nation, determine overall performance, and identify areas of strength and need.

TerraNova

Form/Level: B-14
Test Date: 2/15/99 Scoring: PATTERN (IRT)
QM: 23 Norms Date: 1996

District: Any District

City/State: ANYTOWN, WISCONSIN

CTB McGraw-Hill *Page 2*

Students	Scores	Reading	Language	Math	Science	Social Studies	Writing Prompt: Informative
School Summary Ⓐ	NPNCE	58	52	63	70	61	-
	MNS	5.4	5.1	5.6	6.0	5.6	-
	MNCE	54.6	51.1	57.0	60.8	55.8	-
	MSS	647.8	640.8	636.9	652.2	645.8	-
	MHOLS						3.6
Number of Students		70	70	71	71	71	71
Ⓑ							
Proficiency Level 1 Minimal Performance	PCT	8	3	3	0	1	
	NUM	6	2	2	0	1	
Proficiency Level 2 Basic	PCT	17	39	25	11	19	
	NUM	12	28	18	8	14	
Proficiency Level 3 Proficient	PCT	57	39	51	61	58	
	NUM	41	28	37	44	42	
Proficiency Level 4 Advanced	PCT	15	17	19	26	19	
	NUM	11	12	14	19	14	
Total Enrollment: 72 (See note below)							

Group (Summary) Scores
NPNCE: NP of the Mean NCE
MNS: Mean National Stanine
MNCE: Mean Normal Curve Equivalent
MSS: Mean Scale Score
MHOLS: Mean Holistic Score (Writing)

Note: Percentages are based on total enrollment, including students for whom no scores are reported, and may not total 100%.

PCT: Percent at Level
NUM: Number at Level

Copyright © 1997 CTB/McGraw-Hill

FIGURE 3.7
School Record Sheet *Continued*

From *Wisconsin Student Assessment System* (based on the Terra/Nova tests), 1997, Monterey, California: CTB/McGraw-Hill. Copyright 1997 by CTB/McGraw-Hill. Reproduced with permission of CTB/McGraw-Hill.

For example, let's examine the first row of scores.

- The first column tells you the scores to the right are for a student born on January 16, 1985, who took Form B Level 18 of the WKCE.
- The third column tells you the student's scores in Reading were:

  NP (National Percentile): 88.

  NS (National Stanine): 7.

  NCE (Normal Curve Equivalent): 75.

  SS (Scale Score): 716.

  GE (Grade Equivalent): 8.5.

  PL (Proficiency Level): 3 (Proficient).
- The last column tells us the student's response to the Informative Writing Prompt earned a 4.5 (i.e., one rater scored it a 4, and the other scored it a 5).

> Let's look at a second example. Look at the fifth student's (Birth date 05/28/85) scores in Math. This student is above average relative to the national percentile (NP = 78) and consequently has a grade equivalent higher than average (8.9). However, the student is not proficient in math (Proficiency = Basic). This shows the difference between grade equivalents, which are set to the norm group, and proficiency levels, which are set to curricular standards for mastery. It is possible to be above average and still not be proficient.

Finally, page 2 of the School Record Sheet reports summary data for a fourth-grade class. The top section (letter A) presents the average scores for the class, and the bottom section (letter B) presents the number and proportion of students in each proficiency level.

The first row presents the averages for fourth-grade students who took the WKCE. The scores reported for academic subject matter areas are:

- **NPNCE** (National Percentile of the average NCE; range 1–99).
- **MNS** (Mean National Stanine; range 1.0–9.0).
- **MNCE** (Mean Normal Curve Equivalent; range 1.0–99.0).
- **MSS** (Mean Scale Score; range 450.0–899.0).
- **MHOLS** (Mean Holistic Score; range 1.0–6.0).
- The last line is the number of students who took the test.

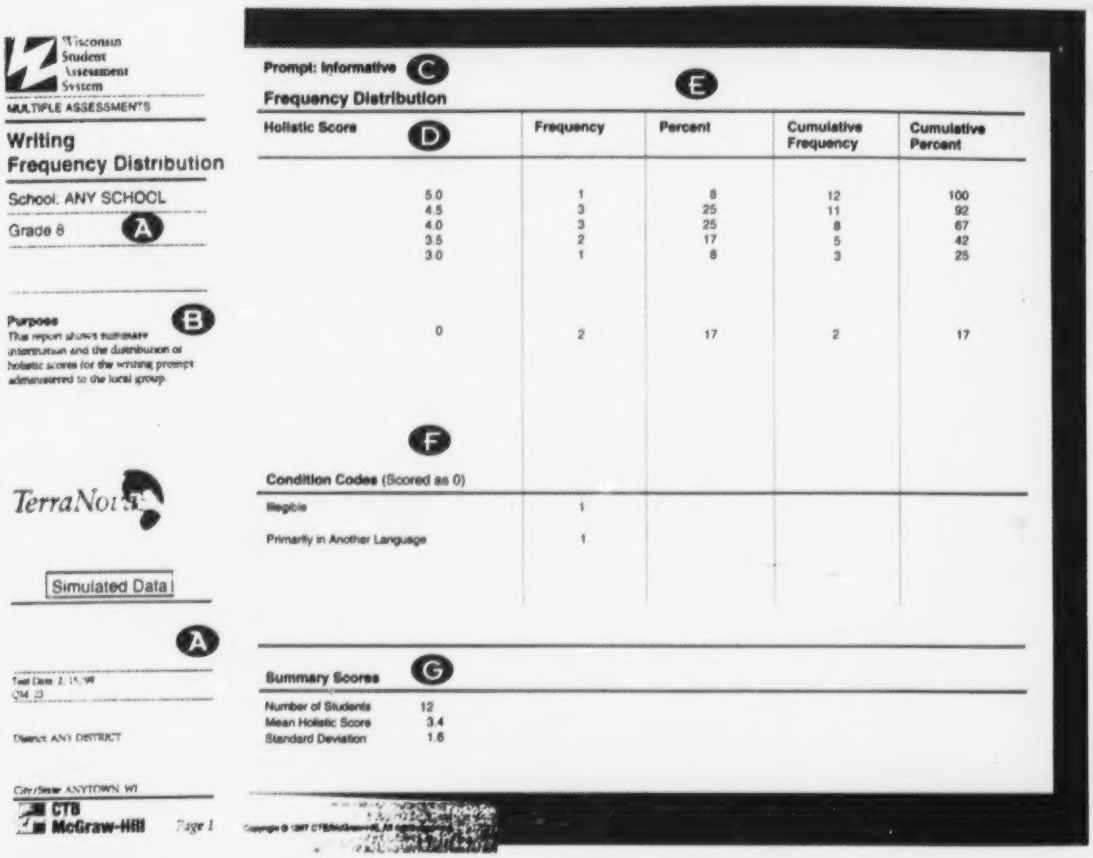
For example, the information in the Reading column tells you that the NP for the average NCE is 59, the average stanine is 5.4, the average NCE is 54.6, the average scale score is 647.8, and 70 students took the Reading WKCE. No average holistic score is reported for reading, because only the writing sample receives a holistic score. The average holistic score for the responses to the Informative Writing Prompt was 3.6.

Continuing our example, the bottom part of the page reports the percentage (PCT) and number (NUM) of students in each proficiency level. Looking at Language, 3% (or 2 students) placed in the Minimal Performance level, whereas 19% (14 students) placed in the Advanced level on the Social Studies test. Note that the percentages are based on the number of students enrolled (72), not the number of students who took the WKCE (70 or 71). Consequently, none of the percentages adds to 100%

Writing Frequency Distribution

Figure 3.8 presents the Writing Frequency Distribution Report for a class of 12 eighth graders. The first column (letter D) lists the scores obtained by class members. The second column shows the number (Frequency) of students who obtained each score. The third column converts the number to the Percent of the class receiving each score. The fourth column converts the number to a Cumulative Frequency (i.e., the number of students in that category plus the number below that category) and the fifth column converts the cumulative total to the Cumulative Percent.

In the example in Figure 3.8, the second column tells you that 2 students received a holistic score of 0; 1 received a score of 3.0; 2 received scores of 3.5;


A
B
C
D
E
F
A
G
TerraNova

Wisconsin Student Assessment System
MULTIPLE ASSESSMENTS

Writing Frequency Distribution

School: ANY SCHOOL

Grade 8

Purpose
This report shows summary information and the distribution of holistic scores for the writing prompt administered to the local group.

Simulated Data

Test Date 2/15/99
QM 23

District ANY DISTRICT

City/State ANYTOWN, WI

CTB McGraw-Hill Page 1

Prompt: Informative

Frequency Distribution

Holistic Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
5.0	1	8	12	100
4.5	3	25	11	92
4.0	3	25	8	67
3.5	2	17	5	42
3.0	1	8	3	25
0	2	17	2	17

Condition Codes (Scored as 0)				
Illegible	1			
Primarily in Another Language	1			

Summary Scores

Number of Students	12
Mean Holistic Score	3.4
Standard Deviation	1.6

Copyright © 1997 CTB/McGraw-Hill. All [illegible]

FIGURE 3.8
Writing Frequency Distribution

From *Wisconsin Student Assessment System* (based on the Terra/Nova tests), 1997, Monterey, California: CTB/McGraw-Hill. Copyright 1997 by CTB/McGraw-Hill. Reproduced with permission of CTB/McGraw-Hill.

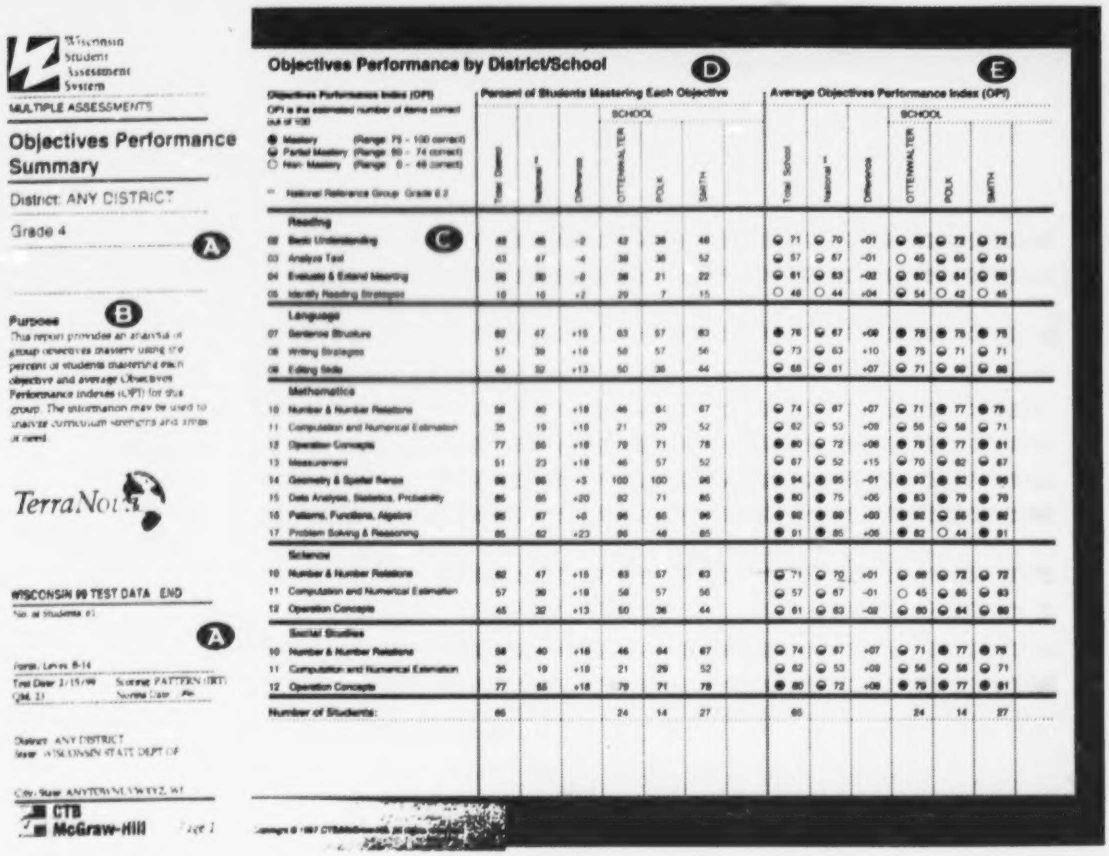
3 received scores of 4.0, 3 received scores of 4.5, and 1 received a score of 5.0. No students received scores of 1–2.5, and no student scored above 5.0 in this classroom.

The third column shows that a quarter of the class (3/12 = 25%) earned a score of 4.0, but only 8% of the class (1/12 = 8%) earned scores of 3.0 and 5.0. The fourth column shows that 8 students had scores of 4.0 or less. The last column shows that 25% of the class (3/12 = 25%) had scores of 3.0 or less, and 67% (8/12) had scores of 4.0 or less.

The rows in the Condition Codes section (letter F) explain why two students earned scores of 0 (remember, the lowest Holistic Score is 1). One student's response was illegible, and one was written primarily in another language. None were off topic or insufficient (i.e., too short) to evaluate.

The Summary Scores at the bottom of the report (letter G) tell you that 12 students took the exam. The mean holistic rating for these 12 students is 3.4, and the average spread (standard deviation) of holistic ratings is 1.6.

BEST COPY AVAILABLE



Wisconsin Student Assessment System

MULTIPLE ASSESSMENTS

Objectives Performance Summary

District: ANY DISTRICT

Grade 4

Objectives Performance by District/School

A
B
C
D
E
TerraNova
WISCONSIN 99 TEST DATA END
CTB McGraw-Hill

FIGURE 3.9
Objectives Performance Summary

From *Wisconsin Student Assessment System* (based on the Terra/Nova tests), 1997, Monterey, California: CTB/McGraw-Hill. Copyright 1997 by CTB/McGraw-Hill. Reproduced with permission of CTB/McGraw-Hill.

Objectives Performance Summary

This section provides information about how well the class performs within specific academic skill objectives (i.e., academic objectives). Figure 3.9 reports the outcomes for a district's fourth grade of 65 students taking the WKCE. The major row divisions present specific academic objectives in five subject domains (Reading, Language, Mathematics, Science, and Social Studies). There are two groups of columns. The left-most group of six columns presents information about the percentage of the class whose Objective Performance Index (OPI) score is greater than or equal to 75% (i.e., the percentage of students whom you might assume have mastered the objective). The right-most group of six columns describes the average, or mean, OPIs for the class by academic objective.

Within each of these divisions (left and right), there are six columns. Each of the columns presents information as follows:

BEST COPY AVAILABLE

- Total District (information for the entire district. Note that the first column on the right side of the page is mislabeled "Total School"; it should read "Total District").
- National (the national average).
- Difference (the difference between the district minus the national average—negative numbers imply the district is below the national average; positive numbers imply the district is above the national average).
- OTTENWALTER (name of first school in the district).*
- POLK (name of second school in the district).*
- SMITH (name of third school in the district).*

An example will help you understand what these numbers mean. First, look at the first row. It summarizes information for the grades of 65 students regarding four Reading objectives (Basic Understanding, Analyze Text, Evaluate and Extend Meaning, and Identify Reading Strategies). Look at the first line in this column; it reports information about how well students did on the academic objective of Basic Understanding.

- The first column to the right reports the outcomes for all the fourth graders in the district (Total District). So, the first number (43) means that 43% of the fourth graders in this district earned an OPI of at least 75%. Another way of saying this is that you might guess 43% of the students in this grade have mastered Basic Understanding skills in reading.
- The second column (National) presents the proportion of students in the national sample who earned OPIs of 75% or greater. In this example, the number is 45, meaning in a typical classroom, you might expect 45% of the students to have mastered Basic Understanding.
- The third column (Difference) reports the difference between the Total District and National columns. In this case, the number (–2) means the percentage of fourth graders in this school who have mastered Basic Understanding skills in Reading is slightly less (by 2%) than the proportion of the national sample who have mastered these skills. When a district performs better than the national average, the numbers in the Difference column will be positive; when the district performs worse, the numbers will be negative.
- The fourth through sixth columns report the percentage of students at each school in the district who have mastered each objective. So, the percentage of fourth graders at Ottenwalter who have mastered Basic Understanding is 42%, whereas only 36% of fourth graders at Polk have mastered the skill.

---

* The number of school columns may vary from 1 (repeats data for the entire district) to as many classrooms or schools as the district wants to report.

The columns on the right side of the report present the mean, or average, OPI for the class. Means are presented in two ways: visual symbols reflecting three levels of achievement (nonmastery, partial mastery, and mastery), and the actual number of the mean. Look at the top line of results to see how students did for Basic Understanding in Reading.

- The first column in this section (Total School) reports the mean OPI for all fourth graders in the district. The number 71 means the average OPI was 71; because 71 is between 50 and 74, it falls in the partial mastery range (50 and 74), (see upper left corner of the report for a key).
- The second column in this section reports the mean OPI for the national sample of fourth graders. The number 70 means the average for the national sample was 70, which falls in the partial mastery range of 50 and 74, and so is illustrated with a half-filled circle .
- The third column (Difference) reports the difference between the mean OPI for the Total School (really, district) and the mean OPI for the national sample. In this case, 71–70 = +01. In other words, the average OPI for this school's fourth grade was higher (by +1%) than the national average OPI.
- The fourth through sixth columns in this section report the mean OPI for each fourth grade at each school in the district. The average OPI for Basic Understanding at Polk was higher (72) than the average for Ottenwalter (69).

Examination of the Objectives Performance Summary is probably the most useful activity for planning instruction. You might look down through the first (left-most) column to find objectives students have mastered (i.e., those you have successfully taught) and those that students have not mastered (i.e., those you have not successfully taught). For example, the high proportion of students mastering Mathematics objectives suggests that these are strong areas of instruction. However, within this instructional domain, student mastery of Computation and Numerical Estimation is relatively low. By examining the proportion of a class that has mastered objectives, or by examining objectives with relatively high and low mean OPIs, you can identify areas of strength and areas in need of improvement within your instruction.

Note: the example in Figure 3.9 lists the same three objectives (10, 11, 12) for Mathematics, Science, and Social Studies. This is an error. The Science and Social Studies objectives are identified incorrectly.

Please keep in mind two important points when using OPIs to shape your teaching. First, OPIs are not the same as proficiency levels. OPIs are linked to specific academic objectives, not general academic proficiency. Their specificity can help you focus your teaching by suggesting relatively weak or strong areas of instruction within academic domains. However, proficiency levels reflect an aggregate performance within a broader domain.

Second, always validate the results of standardized tests with your own assessments. That is, check the results of tests against student work, quizzes, exams, and other evidence of student performance you collect in your classroom. Often, teachers do not teach, or test, the academic skills on which students do poorly. For example, the results in Figure 3.9 might suggest that the teacher's approach to mathematics may overlook or fail to provide sufficient practice in computation and estimation. You may want to align instructional content with assessment. However, if you find that results of standardized tests conflict with results of classroom tests (e.g., students do well on your exams but not on the LSA used in your state or district), examine how you ask students to perform versus how the standardized examinations ask students to perform. You may find it useful to align your assessment methods to those of your state or district LSA. (See Figure 2.4, page 26, for guidance on how to use assessment results for instructional decisionmaking.)

Summary

As you can see, most state and district LSA provides an amazing amount of information on individual students, classrooms, and districts. However, it may be that LSA can provide too much of a good thing. Because many educators are uncertain about how to interpret LSA scores, and the meaning of individual and aggregate results, they may feel overwhelmed by LSA results. Understandably, many educators may simply choose to ignore or dismiss the results in favor of evidence that they find more relevant and appropriate to their daily activities (e.g., the responses of students on quizzes, oral statements in class, answers to questions). This is unfortunate, because we believe that periodic LSA data complement more regular and informal measures of student learning. That is, LSA data can inform and invite educators to think strategically, whereas everyday interactions, tests, quizzes, and papers invite educators to select specific tactics and methods. Understanding LSA can add a strategic element to educators' instructional planning and teaching activities.

Applying Your Knowledge of LSA: Case Applications

Now that you have read about LSA, it's time to revisit our case reflections and apply our knowledge of LSA to the cases. Note that you must know about the content and structure of the LSA used in your district or state to apply your knowledge of LSA to the students you serve. Look at the cases on the next page for information.

Commonly Asked Questions and Answers About LSA

This chapter is intended to enhance your assessment literacy for understanding and interpreting LSA results. The first part of the chapter outlined why assessment literacy is important to teachers. The second part of the chapter described the Wisconsin Knowledge and Concepts Exam content and results, and the third part of the chapter provided opportunities for you to apply your knowledge of the WKCE to interpreting results. However, you still may have some questions about the examinations. We often have been asked some of the following questions.

Case Applications and Large-Scale Assessment

Patrick

■ Patrick will take the Florida Comprehensive Assessment Test (FCAT). The FCAT is based on the Stanford Achievement Test—9th Edition (SAT-9). The content and structure of the FCAT are described on Florida's Department of Education website (http://www.firn.edu/doe/). Florida has standards in English/Language Arts, Mathematics, Social Studies, and Science. The FCAT system includes multiple-choice and short-answer items in all subject matter areas and an extended English/Language Arts response. Test performance is reported as criterion-referenced scores (using proficiency levels 1–5; see Table 3.3) and as general norm-referenced scores. Florida has recently passed a law that requires students to pass tests to be eligible for promotion to higher grades and also for the purpose of graduation. Thus, there are some important consequences of not doing well on large-scale assessments in Florida.

Tia

■ Tia will take the Wisconsin Knowledge and Concepts Examinations (WKCE). The WKCE is based on the TerraNova. The content and structure of the WKCE are described on Wisconsin's Department of Education website (http://www.dpi.state.wi.us/dpi/). Wisconsin has standards in English/Language Arts, Mathematics, Social Studies, and Science. The WKCE includes multiple-choice and short-answer items in all subject matter areas and an extended English/Language Arts response. Test performance is reported as criterion-referenced scores (using four proficiency levels; see Table 3.3) and using multiple norm-referenced scores. We have provided extended examples of Wisconsin's reports to illustrate how to interpret test results. Wisconsin recently passed a law that requires students to pass tests for promotion to higher grades and to graduate from high school. Thus, testing for Tia has some high stakes.

Chris

■ Because Chris's parents expressed interest in having Chris take the state test, Chris's teachers have begun to examine the test in greater detail. They were not familiar with the test, because they had typically not worked with students who took these tests. Although their conclusions will be addressed in later chapters, Chris's home state (Idaho) uses a mix of standardized tests (the Iowa Tests of Basic Skills) and extended responses scored by rubrics. Idaho's website (http://www.sde.state.id.US/Dept/) provides information about its standards and assessment system. Idaho has standards in all core subject matter areas, and reports results using norm-referenced and criterion-referenced scores (see Table 3.3). Idaho also has an alternate assessment for students who require significant amounts of instructional support and whose curriculum focuses on functional living skills. Thus, once Chris's teachers learn more about the content of the Iowa Tests of Basic Skills, they will be armed with the information they need to determine which aspect of the Idaho assessment system Chris will be involved in taking.

1. **How well does the content of state tests align with state academic standards?**

   The answer varies state by state, but some general conclusions hold across most states. First, most tests sample only about half of a state's academic standards, but they rarely sample content not found in the standards. In other words, the examinations are essentially free from irrelevant academic skills and content, but they are incomplete. Some entire domains, such as oral communication, music, and physical education, are not included in most state LSAs, and neither are some parts of some domains. Second, most tests sample content at a less complex level than the standards define. Most standards call for complex activities such as problem solving, reasoning, and application of content knowledge, yet LSA rarely taps these processes as deeply as demanded by state standards (Linn, 2000).

2. **How well does LSA measure what students learn in a classroom?**

   It depends on the degree of alignment between classroom instructional activities and exam content. If the classroom's curriculum and instruction are closely aligned to test content, the exams will provide a good measure of student learning. However, if the classroom's curriculum and instruction are poorly aligned to exam content, the exam will not reflect student learning. Usually, states encourage teachers to align their instructional content to state standards, not state tests. Because tests are aligned to standards, and instruction is aligned to the same standards, the tests ought to provide a reasonable indicator of student learning.

3. **How is curriculum alignment different from "teaching to the test"? Isn't it wrong to "teach to the test"?**

   Aligning curriculum, instruction, and assessment is essential to effective education, but teaching to the test is cheating. How should you separate CIA alignment from teaching to the test? The answer is in the specificity of the teaching. If you teach to the instructional objectives sampled by LSA, and assess student progress by requesting similar kinds of responses, you are aligning curriculum, instruction, and assessment. If you teach the answers to a specific set of items or questions you think might be on the test, you are teaching to the test. This difference might be illustrated by a driving instructor who taught students only the exact sequence of driving activities needed to pass a behind-the-wheel test at a specific motor vehicle office (teaching to the test), versus an instructor who taught the elements of driving that might be included on the test (aligning instruction). Alignment promotes knowledge and skills students can use regardless of specific item content; teaching to the test promotes knowledge and skills that are useful only for a specific set of items.

4. **What is the reading level of most LSAs? Aren't grade-level tests useful only for students who are on grade level?**

   Most tests used in LSAs span a range of grade levels. For example, a typical fourth-grade test will have a readability range from approximately beginning second-grade to fourth-grade level. A typical eighth-grade test has a readability range from early fourth grade to eighth grade. Likewise, test content spans a range of usually two (or more) grade levels on either side of the grade in which it is given. Thus, most tests will allow even students substantially below grade level to participate effectively in the test.

5. **What do test publishers say about the use of testing accommodations with their tests?**
   Most test publishers do not take a position on the use of testing accommodations. It is also rare for publishers to include accommodations in the development of their tests. Consequently, publishers rarely provide users with specific guidance about appropriate or inappropriate testing accommodations.
6. **The examination appears to measure knowledge, skills, and the application of these within subject matter areas, but does little to assess integration of skills across subject areas such as mathematics and science or language arts and social studies. Why? This is inconsistent with efforts to provide students with integrated curriculum and instruction.**
   Most tests are designed to focus on knowledge, skills, and their application primarily within core subject matter areas of reading/language arts, mathematics, science, and social studies because that is how state content and performance standards define learning objectives. This approach maximizes the ability to isolate academic skills within subject matter, but it minimizes the understanding of integrated subject matter knowledge.
7. **Are there practice materials or recent past tests available so teachers and students can get a clear understanding of the types of questions asked on the test and the array of item formats or types?**
   Yes. First, test publishers provide practice activities. Typically, these are booklets with five or six practice items in each of the core subject matter areas. The items and the test directions are representative of those on recent versions of the test. Second, you may be able to review a copy of last year's examination by contacting your school assessment coordinator. Copies of the forthcoming year's examination are usually secure until after the test is given and the test response forms are returned for scoring. Third, most states offer sample items and other materials to explain test content and formats on their websites and through publications.
8. **Are tests available in other languages for students with limited English proficiency?**
   The major achievement tests are available in Spanish language editions. However, many states do not allow the Spanish version to substitute for the English version of the test, because the scores have not been demonstrated to be equivalent, and because state standards often call for proficiency in academic knowledge and skills in English.

CHAPTER FOUR

Inclusive Assessment Tactics: Testing Accommodations and Alternate Assessments

Public schools in the United States serve more than 60 million students, all of whom are expected to learn and progress toward productive lives as citizens. Included in this population of students are more than 6 million students identified with disabilities. All of these students with special needs have individualized education programs (IEPs) or individualized accommodation plans (IAPs) developed with input from parents and educational specialists. The majority of these students have relatively mild disabilities and, in most cases, are being taught much of the same content as their peers without disabilities, but possibly with different instructional methods or different developmental time lines.

Documenting the achievements and educational progress of students is a critical aspect of an appropriate education and is required by law for students with disabilities. Consequently, educators are responsible for collecting evidence that students are learning. As we have emphasized in the previous chapters, state and district assessment programs featuring achievement tests are one of the primary methods educators use to collect evidence of students' learning. Typically, when educators think of testing students with disabilities, however, they usually think about individualized, norm-referenced tests of cognitive abilities, achievement, and social and adaptive behavior used to identify students who may have disabilities and have special educational needs. Such tests are often helpful in identifying students with disabilities, but they provide limited evidence concerning educational progress, because they usually do not contain specific content that is aligned with what most students are being taught daily. In addition, such tests' scores are usually designed to be interpreted as norm-referenced scores and do not allow for progress comparisons from time 1 to time 2, to other students in the same schools, or to established proficiency standards.

In communities across the nation, many educational stakeholders want educators to be more accountable and to emphasize high standards for all students. Currently, 48 states have statewide tests for the purpose of monitoring students' educational achievement. (Only Iowa and Nebraska do not have such testing programs, although school districts in these states do a considerable amount of achievement testing.) The vast majority of state assessment programs and other tests used by school districts have been and will most likely continue to be part of the evidence used to document what students are learning and how well they are learning it.

All Means ALL

Historically, not all students have been included in many of the statewide or schoolwide assessment efforts. Participation rates for students during the past several years in statewide assessments have ranged from a low of 33% to a high of 97% (Thurlow, Nelson, Teelucksingh, & Ysseldyke, 2000). Many of the students who did not participate were students with disabilities or with limited English proficiency. Why has this happened? There are several possible reasons for these varying participation rates. In most states it means that thousands of students' achievement has gone unmeasured. However, if educators and other educational stakeholders who aspire to high standards for all students are to have a meaningful picture of how well students are learning and applying valued content knowledge and skills, all students need to be assessed periodically. The cases of at least two of our students, Tia and Chris, may provide some suggestions as to why and how students with disabilities have often been excused or excluded from large-scale assessments (See Case Reflections Box). In some states, even students like Patrick, who has difficulty reading but has not been classified as a student with a disability, may not be participating in the state's assessment program.

Reasons typically given for excluding students like Patrick, Tia, and Chris from testing programs include the following:

- The concern that students with disabilities or with significant reading difficulties will lower a school's mean score.
- The desire to "protect" students with disabilities from another frustrating testing experience.
- The perception that the tests are not relevant, especially to students with disabilities.
- The fact that some parents do not want their child spending valuable class time taking a test that doesn't count toward a grade.
- The belief that the guidelines for administering standardized tests prohibit, or at least greatly limit, what can be changed without jeopardizing the validity of the resulting test score.

The limited participation of students with disabilities in state and district assessments results in

- Unrepresentative mean scores and norm distributions.
- Incomplete picture of student performance related to educational standards.
- Beliefs that students with disabilities cannot do challenging work.
- The undermining of inclusion efforts for many students.
- Possible legal sanctions that could mean the loss of significant amounts of federal financial support.

Case Reflections and Accommodations

Patrick

■ Patrick is youngster whose reading skills have developed slowly. Consequently, he has struggled somewhat in school, but he does not qualify as a student with a disability. What, if anything, can be done to help him demonstrate on Florida's statewide assessment what he knows and can do in subject matter areas such as mathematics, social studies, and science? Without some form of accommodation, his reading difficulties are sure to interfere with his performance in all subject matter areas.

Tia

■ Tia works hard in class and is eager to be part of her peer group. Her reading disability clearly interferes with her comprehension of written material like that which is on her state's test. What, if anything, can be done to facilitate her meaningful participation in Wisconsin's statewide assessment system? How is she likely to feel about receiving permissible accommodations when most of her classmates will not be able to have accommodations when they take the same test?

Chris

■ Given the severity of Chris's disability and the fact that he receives all of his instruction in a special education classroom where functional living skills are emphasized, it is very unlikely that he could participate in Idaho's large-scale achievement test without significant accommodations. But even with accommodations, the results of the test may not be meaningful or valid. What alternatives are there for including Chris in the statewide accountability system if he does not take the same test as other students in the state?

? What can be done legitimately to help each of these students to participate in their state's assessment? What accommodations are likely to be permissible and useful? Could they qualify for an alternate assessment if accommodations are deemed ineffective?

Since the passage of federal and state legislation in the 1970s, students with disabilities have been guaranteed access to a free, appropriate public education. Therefore, when tests and assessment systems are designed to serve as indicators of progress in the subject matter content of a school's curriculum or the state's academic standards and are used to make decisions about future educational services, all students are expected to participate in the assessments as part of their free, appropriate public education. The legal basis for this position is based on a number of federal laws, including Section 504 of the Rehabilitation Act of 1973, Title I of the Elementary and Secondary Education Act (ESEA), the Americans with Disabilities Act of 1990 (ADA), and, most recently with regard to children with disabilities, in the 1997 amendments to the Individuals with Disabilities Education Act (IDEA '97, Public Law 105-17). Assessment is often associated with direct individual benefits such as promotion, graduation, and access to educational services. In addition, if done well (as discussed in Chapters 2 and 3), assessment is an integral aspect of educational accountability systems. Such assessments provide information that could benefit individual students by measuring individual progress against standards or by evaluating programs. Because of the potential benefits that may occur as the result of assessment, exclusion from such assessments on the basis of disability generally would be a violation of Section 504 and ADA.

For More Information on

IDEA, see:

http://www.ideapractices.org

Title I and IDEA include a number of specific requirements for including all children in assessments. Heumann and Warlick (2000) noted that in adding these requirements, Congress recognized that many students were at risk (i.e., students with disabilities, minority children, children with limited English proficiency) and were not experiencing levels of achievement in school that would enable them to successfully pursue postsecondary education or competitive work opportunities. Many of these children's educational programs were characterized by low expectations, limited accountability for results, and exposure to poorer curricula than offered to other children. Thus in the 1997 amendments to IDEA, there are requirements concerning the following:

- The participation of children with disabilities in general statewide and districtwide assessment programs, with appropriate accommodations, when necessary.
- Documenting in a student's IEP any individual modifications in the administration of state or district tests that measure achievement.
- Documenting in a student's IEP a justification for exclusion from a standardized test and indicating how the student will be assessed with an alternate method.
- Reports to the public about the participation and performance of children with disabilities with the same details as reports for children without disabilities.

Making decisions about including students with disabilities in assessment programs and validly implementing assessments requires teachers' active involvement on IEP teams and can be challenging. One of the first challenges confronting educators is to determine the "right" assessment program for students with disabilities. Practically speaking, students with disabilities could participate in (a) the regular assessment without accommodations, (b) the regular assessment with testing accommodations, or (c) an alternate assessment. And in some states a fourth option exists whereby a student could participate in part of the regular assessment with testing accommodations and the remainder in an alternate assessment. In making this participation decision, educators consider an array of factors, many of which are "magnified" in Figure 4.1. As highlighted in this figure, several of the most critical factors include (a) the alignment between a student's IEP goals, classroom curriculum, and the content of the test; (b) a student's reading ability; and (c) the nature of instructional accommodations a student typically receives.

Much more will be said about making participation decisions after we examine two methods or tactics that can be used to facilitate the participation of students who traditionally have been left out, excused, or exempted from large-scale assessments. These tactics are testing accommodations and alternate assessments.

Tactics for Increasing the Meaningful Participation of *All* Students in Assessment Programs

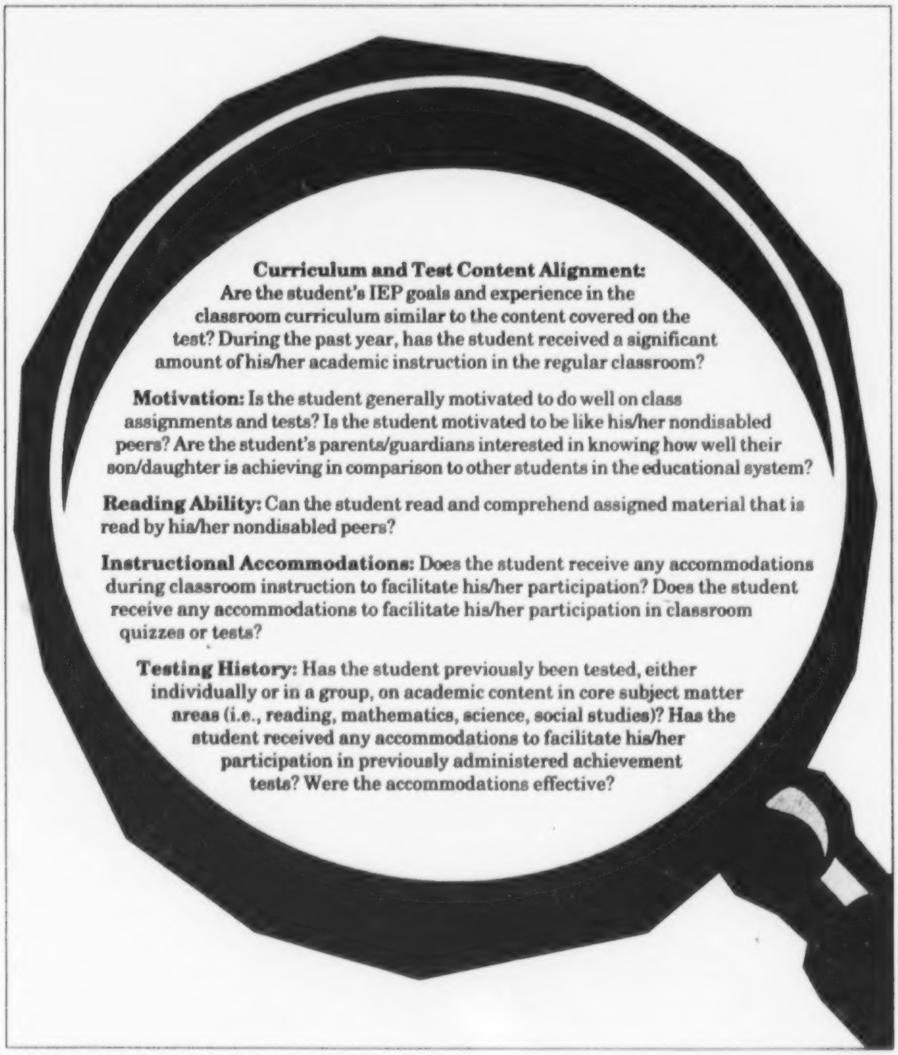
As noted in IDEA '97 and in many states' guidelines or testing policies, testing accommodations and alternate assessment are two methods that educators can use to facilitate the participation of all students with disabilities in assessments and accountability systems. Therefore, every teacher who works with students with disabilities should know about testing accommodations and alternate assessment if they want to facilitate their students' meaningful involvement in assessment programs.

Testing Accommodations

One of the most frequent steps for increasing the meaningful participation of students with disabilities in assessments is allowing changes to testing procedures. Such changes are commonly referred to as *testing accommodations*, although as noted in a recent National Center for Educational Outcomes (NCEO) Policy document (National Center for Educational Outcomes, 2000) several states and the widely respected *Standards for Educational and Psychological Testing* (American Educational Research Association, 1999) use the term *modification* instead of or interchangeably with the term *accommodation*.

Definition and Purposes

Testing accommodations are *changes in the way a test is administered or responded to by a student*. Testing accommodations are intended to offset distortions in test scores caused by a disability without *invalidating* or changing what the test measures (McDonnell et al., 1997). Many different testing accommodations are



Curriculum and Test Content Alignment: Are the student's IEP goals and experience in the classroom curriculum similar to the content covered on the test? During the past year, has the student received a significant amount of his/her academic instruction in the regular classroom?

Motivation: Is the student generally motivated to do well on class assignments and tests? Is the student motivated to be like his/her nondisabled peers? Are the student's parents/guardians interested in knowing how well their son/daughter is achieving in comparison to other students in the educational system?

Reading Ability: Can the student read and comprehend assigned material that is read by his/her nondisabled peers?

Instructional Accommodations: Does the student receive any accommodations during classroom instruction to facilitate his/her participation? Does the student receive any accommodations to facilitate his/her participation in classroom quizzes or tests?

Testing History: Has the student previously been tested, either individually or in a group, on academic content in core subject matter areas (i.e., reading, mathematics, science, social studies)? Has the student received any accommodations to facilitate his/her participation in previously administered achievement tests? Were the accommodations effective?

FIGURE 4.1
"Magnifying" Key Variables Discussed by IEP Teams When Making Participation Decisions and Accommodations

Note. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 85. Copyright February 2000 Wisconsin Department of Public Instruction.

allowable as long as they do not reduce the validity of the test scores. In virtually all states, the IEP team is given guidance by state testing guidelines or policies but ultimately is entrusted to determine the appropriate testing accommodations for individual students with disabilities. (For details on testing accommodations that are generally acceptable in your state consult your state's testing accommodations policies, the *NCEO Synthesis Report 33 on State Participation and Accommodation Policies for Students with Disabilities: 1999 Update,* or visit the NCEO website at www.coled.umn.edu/NCEO).

BEST COPY AVAILABLE

For More Information on
States' Testing Accommodations, see:
http://www.coled.umn.edu/NCEO

Tests and assessment programs can be altered in a variety of ways to facilitate the participation of students with disabilities and provide valid results. As increasing numbers of students with disabilities are included in assessment programs and take the same tests as their peers without disabilities, it is essential that teachers and other members of IEP teams consider the use of testing accommodations. It is important to understand that accommodations are intended to maintain and facilitate the measurement goals of an assessment, not to modify the actual questions or content of the tests. Accommodations usually involve changes to the testing environment (e.g., braille or large-print materials, the amount of time a student has to respond, the quietness of the testing room, assistance in reading instructions) or the method by which a student responds to questions (e.g., orally with a scribe, pointing to correct answers). Testing accommodations should not involve changes in the content of test items. When this occurs, the test is very likely to be measuring different skills or different levels of the same skills, and consequently the conclusions made from the results are likely to be invalid.

Accommodations generally result in some minor changes in the procedures for administration or response upon which a test was standardized. Because many educators have been taught to follow standard administration procedures exactly, there may be some reluctance to use accommodations. There are three keys to the selection and appropriate use of testing accommodations. First, accommodations must be determined on a *case-by-case basis* for each student in each subject tested. Second, knowledge of the instructional accommodations a student currently receives should guide considerations of testing accommodations. Third, accommodations are intended to make the test a more accurate measure of what a student knows or can do. That is, IEP teams must select accommodations that are likely to facilitate a student's participation in a testing program but not likely to change or invalidate the intended meaning of a test score. In effect, testing accommodations are intended to increase the validity of a student's test score.

To date, there is not a comprehensive research base to guide educators' decisions about which accommodations invalidate test results or which accommodations improve test performance without invalidating test results. Studies of the effects of testing accommodations on test scores of students with disabilities have been published, and numerous investigations are underway in major research centers and test companies across the country. However, decisions about testing accommodations and their effect on a student's test performance are highly individualized events, so research on testing accommodations is unlikely to be prescriptive enough to satisfy most educators. Nevertheless, any individual potentially involved in making important decisions should have a general understanding of what researchers examining testing accommodations have learned. Therefore, in a later section of this chapter we provide a current "snapshot summary" of some of the best research on testing accommodations.

Even if you haven't kept up to date with the current research on testing accommodations, all is not lost. If you have a clear understanding of what a test or subtest measures, many of the decisions about appropriate (i.e., valid) accommodations become rather straightforward. For example, reading questions and answers on a reading test designed to measure sight vocabulary and comprehension would certainly invalidate the resulting score, because these accommodations are changing the skills or competencies that the test is designed to be measuring. Conversely, reading a complex story problem on a test designed to measure mathematics reasoning and calculation could be appropriate for some students with disabilities. In this latter case, assistance with reading is designed to increase the likelihood that the test score is a better indicator of what the student has learned in mathematics. If the accommodation does this, then the test score is said to be valid.

Testing accommodations have commonly been grouped into four categories:

- Accommodations in timing.
- Accommodations to the assessment environment.
- Accommodations in the presentation format.
- Accommodations in the recording or response format.

Figure 4.2 provides some specific examples of each of these categories of accommodations.

It is important to note that not all students with disabilities will need testing accommodations to participate and provide a valid or accurate account of their abilities. On the other hand, for a small number of students with more severe disabilities, testing accommodations will not be appropriate or reasonable. These students' educational goals and daily learning experiences concern content that may differ significantly from that contained in state or district content standards. Although many of the IEP goals of these students should be aligned with the state's academic content standards, a student's current performance may differ significantly from the performance standards expected for the student's grade level. Consequently, students in this situation will need to participate in an alternate assessment to meaningfully measure their abilities and provide valid results.

Many educators find it difficult to make decisions concerning the selection and use of testing accommodations with students. They also find it difficult to explain the use of testing accommodations to other educational stakeholders. We suggest two metaphors for thinking about the role and function of testing accommodations.

The first metaphor for testing accommodations is *eyeglasses*. Look around any room with other adults present and you will see at least one third and maybe one half of them wearing eyeglasses to correct for vision impairments. Eyeglasses are an accommodation for imperfect or poor vision. If you wanted to test the natural vision ability of a person who wears glasses for driving and outdoor activities, then wearing glasses during a test of distance vision would invalidate the test score, assuming that your purpose was to make an inference about the person's natural or uncorrected vision. On the other hand, if your purpose was to determine the same person's driving ability, then wearing the

Time Accommodations

- Administer a test in shorter sessions with more breaks or rest periods.
- Space testing sessions over several days.
- Administer a test at a time most beneficial to a student.
- Allow a student more time to complete the test.

Setting Accommodations

- Administer the test in a small-group or individual session.
- Allow a student to work in a study carrel.
- Place a student in a room or part of a room where he or she is most comfortable.
- Allow a special education teacher or aide to administer the test.

Format Accommodations

- Use an enlarger to facilitate visual perception of material.
- Use a braille transcription of a test.
- Give practice tests or examples before the actual test is administered.
- Assist a student in tracking test items by pointing to or placing the student's finger on the items.
- Allow use of equipment or technology that a student uses for other school work.

Recording Accommodations

- Use an adult to record a student's response.
- Use a computer board, communication board, or tape recorder to record responses.

FIGURE 4.2
Examples of Accommodations Frequently Considered Appropriate for Students with Disabilities

Note. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 85. Copyright February 2000 Wisconsin Department of Public Instruction.

glasses during the driving test that he or she wears daily would be a valid accommodation because it would facilitate a more accurate assessment of the person's driving skills by minimizing or eliminating problems due to vision impairments. Remember, even in the absence of disabilities or other complicating factors, tests are imperfect measures of the constructs they are intended to assess. In summary, testing accommodations are intended to function like a corrective lens that will deflect the distorted array of observed scores back to where they ought to be—that is, back to where they will provide a more valid image of the performance of individuals with disabilities.

The second metaphor for testing accommodations is an *access ramp.* An access ramp can be conceptualized as part of a package of testing accommodations for individuals with significant physical impairments. If individuals can't get to the testing room, then they certainly can't demonstrate what they know or can do! The conceptual value of an access ramp has additional meaning when addressing issues of construct validity. Testing accommodations facilitate access to a test for students with a wide range of disabilities, just as a ramp facilitates access to a building for individuals with physical disabilities. The tests that students are required to take are designed to measure some specific *target cognitive skills or abilities*, such as mathematical reasoning and computations, but they almost always assume that students have the skills to access the test, such as attending to instructions, reading story problems, and writing responses. Thus, knowledge and concepts tests like those included in most state assessment programs target broad constructs such as mathematics, science, social studies, and language arts and are used to determine how students are doing in these subjects. Some students—in particular, many students with disabilities—have difficulty with the *access skills* needed to get "into" the test. (See Box 4.1 for an access skill activity.) Thus, valid testing accommodations, just like an access ramp, should be designed to reduce problems of access to a test and enable students to demonstrate what they know and can do with regard to the skills or abilities the test is targeting.

Research on Testing Accommodations

In 1993, researchers at NCEO published a literature review on testing accommodations for students with disabilities (Thurlow, Ysseldyke, & Silverstein, 1993). They found little published empirical research on testing accommodations and tremendous variability across states in rates of participation of students with disabilities and in testing accommodation guidelines.

Six years later, a comprehensive review of research on testing accommodations, broadly defined, by Tindal and Fuchs (1999) extended the NCEO review but unfortunately still left us well short of a clear understanding of the effects and consequences of testing accommodations. For the purpose of updating you on testing accommodations research, we review four studies of the most relevant studies for K–12 educators. These studies represent efforts to examine the impact of testing accommodations that are individualized according to student needs, which often include accommodations from more than one category (e.g., time, setting, presentation/response formats, assistive devices) and are not merely a standard package of accommodations established by the researcher. Unfortunately, most of the research conducted to date has featured the application of a single accommodation or a standard package of accommodations identified by the researcher rather than applying accommodations based on individual student needs. Given that IDEA '97 stipulates an individualized approach to accommodations, and most state testing guidelines maintain that accommodation decisions are to be made on a case-by-case basis, it seems logical that researchers investigate the effect of individualized accommodation packages rather than isolated or "prepackaged" accommodations.

One study that falls under the category of individualized accommodations is descriptive in nature, but it does provide information to guide future research. Trimble's (1998) report is based on a post hoc analysis of data from the Kentucky Instructional Results Information System (KIRIS) using 4th-, 8th-, and 11th-grade assessments from 1993 to 1996. The KIRIS is an inclusive accountability system. During the assessments, students with disabilities received accommodations commensurate with their instructional accommodations. Accommodations were coded into several categories as a means to begin

Target Skills vs. Access Skills Activity

Test items are designed to measure specific or general skills or abilities. For example, many mathematics items are intended to measure a student's ability to reason, compute, and communicate a solution or result. The skills or abilities that test developers intend the items to measure can be called target skills or abilities. The same mathematics items require a student to attend, read, remember some information, and ultimately respond by bubbling in an answer choice or writing an extended response. These latter skills are generally not what the test developers designed the mathematics items to measure, but without these skills or abilities students cannot access or interact with the test items to demonstrate whether or not they possess the target skills measured by the items. Thus, skills or abilities such as attending, seeing, writing, etc. are considered access skills or abilities. A list of common access skills is provided below. Can you think of additional access skills?

1. Attending
2. Listening
3. Reading*
4. Remembering
5. Writing*
6. Following directions
7. Working by oneself
8. Sitting quietly
9. Turning pages of test booklet
10. Locating test items
11. Locating answer spaces
12. Erasing completely
13. Seeing
14. Processing information in a timely manner
15. Working for a sustained period of time
16. Spelling*

* Some skills such as reading, writing, and spelling are access skills for tests designed to measure mathematics, science, and social studies, but are target skills on most tests designed to measure reading/language arts skills.

Key Premise

Testing accommodations should be designed to only effect deficits in access skills, not target skills. If an accommodation involves one or more of the target skills or abilities a test is designed to measure, it will invalidate the test score.

BOX 4.1

Parts of this boxed information have been adapted from *Wisconsin Student Assessment System* (based on the Terra/Nova tests), 1997, Monterey, California: CTB/McGraw-Hill. Copyright 1997 by CTB/McGraw-Hill. Reproduced with permission of CTB/McGraw-Hill.

Target Skills vs. Access Skills Activity (continued)

Background Information about Subskills Measured on TerraNova

CTB/McGraw-Hill in developing language arts and mathematics for tests like TerraNova uses the following describers to characterize the many subskills their items are designed to measure.

Reading/Language Arts Objectives and Subskills

01 Oral Comprehension
Subskills: literal; interpretive

02 Basic Understanding
Subskills: sentence meaning; vocabulary; stated information; sequence, initial understanding; stated information graphics

03 Analyze Text
Subskills: main idea/theme; supporting evidence; conclusions, cause/effect; compare/contrast; story elements—plot/climax/character/setting, literary techniques; persuasive techniques; nonfiction elements

04 Evaluate and Extend Meaning
Subskills: generalize; fact/opinion; author-purpose/point of view/tone/bias; predict/hypothesize; extend/apply meaning; critical assessment

05 Identify Reading Strategies
Subskills: make connections; apply genre criteria; utilize structure, vocabulary strategies; self-monitor; summarize; synthesize across texts; graphic strategies; formulate questions

06 Introduction to Print
Subskills: environmental print; word analysis; sound/visual recognition

07 Sentence Structure
Subskills: subject/predicate; statement to question; complete/fragment/run-on; sentence combining; nonparallel structure; misplaced modifier; mixed structure problems; sentence structure

08 Writing Strategies
Subskills: topic sentence; sequence; relevance; supporting sentences; connective/transitional words; topic selection; information sources; organize information; writing strategies

09 Editing Skills
Subskills: usage; punctuation; capitalization; proofreading

Mathematics Objectives and Subskills

10 Number and Number Relations
Subskills: counting; read, recognize numbers; compare, order; ordinal numbers; money; fractional part; place value; equivalent forms; ratio, proportion; percent; roots, radicals; absolute value; expanded notation; exponents, scientific notation; number line; identify use in real world; rounding, estimation; number sense; number systems; number properties; factors, multiples, divisibility; odd, even numbers; prime, composite numbers

11 Computation and Numerical Estimation
Subskills: computation; computation in context; estimation; computation with money, recognize when to estimate; determine reasonableness; estimation with money

12 Operation Concepts
Subskills: model problem situation; operation sense; order of operations; permutations, combinations; operation properties

13 Measurement
Subskills: appropriate tool; appropriate unit; nonstandard units; estimate; accuracy, precision; time; calendar; temperature; length, distance; perimeter; area; mass, weight; volume, capacity; circumference; angle measure; rate; scale drawing, map, model; convert measurement units; indirect measurement; use ruler

14 Geometry and Spatial Sense
Subskills: plane figure; solid figure; angles; triangles; parts of circle; point, ray, line, plane; coordinate geometry; parallel, perpendicular; congruence, similarity; Pythagorean theorem; symmetry; transformations; visualization, spatial reasoning; combine/subdivide shapes; use geometric models to solve problems; apply geometric properties; geometric formulas; geometric proofs; use manipulatives; geometric constructions

15 Data Analysis, Statistics and Probability
Subskills: read pictograph, read bar graph; read line graph; read circle graph, read table, chart, diagram; interpret data display; restructure data display; complete/construct data display; select data display; make inferences from data; draw conclusions from data; evaluate conclusions drawn from data; sampling; statistics; probability; use data to solve problems; compare data; describe, evaluate data

16 Patterns, Functions, Algebra
Subskills: missing element; number pattern; geometric pattern; function; variable; expression; equation; inequality; solve linear equation; graph linear equation; solve quadratic equation; graph quadratic equation; model problem situation; system of equations; use algebra to solve problems

17 Problem Solving and Reasoning
Subskills: identify missing/extra information; model problem situation, solution; formulate problem; develop, explain strategy; solve nonroutine problem; evaluate solution; generalize solution; deductive/inductive reasoning; spatial reasoning; proportional reasoning; evaluate conjectures

18 Communication
Subskills: model math situations; relate models to ideas; make conjectures; evaluate ideas; math notation; explain thinking, explain solution process

BOX 4.1 *Continued*

BEST COPY AVAILABLE

Target Skills vs. Access Skills Activity (continued)

Application Activity

Below are several items like those used on tests such as *TerraNova*. Read through each item with the purpose of identifying the *target skills* (the skills the test developers intended to measure) and key *access skills* (skills needed to "get into" the item and to document a response)

Item #1

Target Skills: ____________________

____

____

Access Skills: ____________________

____

____

Choose the sentence that best combines the underlined sentences into one.

The train sped through the tunnel.

The train sped across the bridge.

✔ **A** The train sped through the tunnel and across the bridge.
B The train sped through and across the tunnel and the bridge.
C The train that sped through the tunnel sped across the bridge.
D The train sped through the tunnel and it sped across the bridge.

Item #2

Target Skills: ____________________

____

____

Access Skills: ____________________

____

____

This chart shows the number of different types of fiction books on a bookstore shelf.

mysteries	10
romances	30
historical fiction	30

The bookstore owner put 10 more mysteries on the shelf. Draw a circle graph that shows the fraction of the total number of books for each type of fiction that are now on the shelf. Use the key to label your graph.

■ mysteries
▨ romances
⠿ historical fiction

Please note that all the test items used in this box are examples from *Teacher's Guide to TerraNova* (McGraw-Hill, 1997) and copied with permission.

BOX 4.1 *Continued*

Target Skills vs. Access Skills Activity (continued)

Item #3

Target Skills: ____________________

____

____

Access Skills: ____________________

____

____

Choose the topic sentence that best fits the paragraph.

____________________. *Some of the rain runs off into brooks and streams. Some of it goes into the roots of plants and trees. Some of it even goes back up into the air!*

○ All living things need water.
○ Rain is often collected in tanks.
✓ ○ The rain that falls from the sky is not lost or wasted.
○ Plants that live in the desert have special ways of storing water.

Item #4

Target Skills: ____________________

____

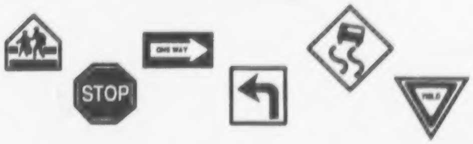
____

Access Skills: ____________________

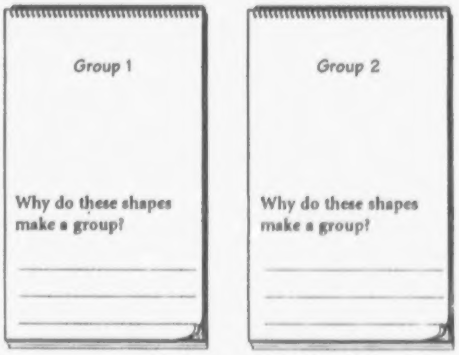
____

____

Look at the shapes of the road signs.


STOP
ONE WAY
YIELD

Sort the shapes into two groups by drawing them on the notepads below. Then explain why the shapes in each group go together.


Group 1
Why do these shapes make a group?
Group 2
Why do these shapes make a group?

Please note that all the test items used in this box are examples from *Teacher's Guide to TerraNova* (McGraw-Hill, 1997) and copied with permission.

BOX 4.1 *Continued*

understanding how certain types of accommodations influence assessment performance. Accommodation categories included: none, reader/oral, scribe/dictation, cueing, paraphrasing, interpreter, technological, and other.

Trimble's (1998) study did not involve experimental manipulation; therefore, no strong conclusions can be reached. However, the results do provide direction for future research. Across all grade levels, students with disabilities performed at a lower level compared to peers without disabilities. The gap between students with and without disabilities was much smaller for 4th grade than for 8th or 11th grade. The combination of accommodations that led to higher student performance varied depending on the grade level, year of test, and content area assessed. For fourth-graders, the following accommodations led to mean performance about equal to that for the total population: oral and dictation combination; paraphrasing, oral, and other combination; dictation; and paraphrasing, dictation, and other combination. In some cases, students with disabilities receiving accommodations, especially those who received the oral accommodation, performed lower than students with disabilities who did not receive accommodations. In both 8th and 11th grades, no combination of accommodations resulted in a higher mean performance for students with disabilities than that of the total population. Overall, the impact of combining the scores of students with disabilities with the general population was marginal (e.g., less than one-tenth of a standard deviation unit).

Trimble (1998) suggested that future research should be experimental. Specifically, he recommended using a comparison group of students without disabilities who receive accommodations to investigate whether accommodations remove barriers due to a disability or unfairly raise performance. Additionally, he suggested testing students under both accommodated and nonaccommodated conditions to compare the impact of accommodations on performance.

Koretz (1997) also conducted post hoc analyses of the KIRIS data. Six specific accommodations were identified as accommodations on the KIRIS: paraphrasing, oral presentation, dictation, cueing, interpreter, and technological aids. Oral presentation, paraphrasing, and dictation were identified as the three most commonly used accommodations. Koretz also reported that the gap in performance between students with and without disabilities was smaller for 4th-grade students than for 8th- or 11th-grade students. However, Koretz reported that the percentage of students with disabilities assessed with accommodations was greater in 4th grade than in 8th or 11th grade. Koretz found that dictation had the strongest effect on scores across grade levels and subject areas. In general, results concerning the impact of accommodations were mixed, depending on subject and grade level. Correlations between item-level performance and total score were similar for all students, regardless of subject, grade, or use of accommodations. However, Koretz found that for students with disabilities who received accommodations, in some instances items were easier, but in other instances items were harder (particularly in the area of math) than for those who did not receive accommodations.

Koretz expressed caution related to the uncertainty of the findings because "it still remains unclear how much the accommodations per se contributed to these disparities" (p. 64). Characteristics of the students, not solely the accommodation(s), may have contributed to the results and given the descriptive rather than experimental nature of the study, alternative hypotheses cannot be ruled out. Of the many recommendations Koretz (1997) made, he emphasized a need for further research on the effects of accommodations that employs different methods than those used in his study to enhance the generalizability of findings to date. Of primary importance is the need for experimental research on the effects of testing accommodations.

Tindal, Heath, Hollenbeck, Almond, and Harniss (1998) conducted an experiment using a large-scale statewide test comparing standard and nonstandard administration procedures featuring two major accommodations (i.e., answering in test booklet vs. bubble sheet, student read vs. teacher read) with fourth-grade students. Participants included 403 students without disabilities and 78 students with disabilities. The reading accommodation was done in a standardized, group manner in which the teacher read each item aloud twice before students answered the question. Tindal and colleagues found no differences between the two response conditions; however, an interaction was found on the math test, indicating that, for students with disabilities, "more valid inferences of math proficiency were possible when students had the test read to them" (Tindal et al., 1998, p. 447). Tindal and colleagues discussed a limitation of the study pertaining to the response condition (i.e., booklet vs. bubble sheet): "As a group, students performed at similar levels in both conditions; however, individuals within the two response conditions may have had higher scores when marking the booklet, but the effect was removed when averaged with other students" (p. 447). This limitation has implications for the design of future studies and, in fact, argues for examining the impact of accommodations on individuals as well as groups of students.

Tindal, Glasgow, Helwig, Hollenbeck, and Heath (1998) also conducted a large-scale study comparing standard and nonstandard test administration procedures for a 30-item multiple-choice math test with students in grades 4, 5, 7, and 8. The accommodations provided included using a videotape to read test items and options, color coding options as each was read, presenting one problem per page, and pacing the test following predetermined solution times for each problem. Results indicated that the videotaped presentation is a viable accommodation that has the potential for improving student performance. The researchers commented, however, that the group design may limit findings in that, even when no significant group gains were noted, they suspected that the accommodation(s) worked for some individual students. They recommended analyzing data using an ideographic approach as a strategy to further understand patterns in the data. Thus, this study also supports the rationale for examining individual as well as group effects.

Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000) used an experimental approach to validate accommodations for use with students on large-scale math tests by administering various curriculum-based measures (CBM) to students with and without disabilities under several types of accommodation conditions (e.g., extended time, read to student, calculator, encoding). They explored group differences—that is, whether students with disabilities experienced a differential boost from the testing accommodations they provided. They also estimated the typical boost that would be expected when they provided accommodations based on performance of students without disabilities. They then compared the accommodation boost of each student with a disability to the typical boost attained by students without disabilities to determine whether a greater-than-expected boost existed for each student with a disability, thereby qualifying for that accommodation on the large-scale assessment. Finally, they compared this method of making accommodation decisions to that of teacher recommendations.

Results of the Fuchs and colleagues (2000) study suggested that, as a group, students with disabilities did not demonstrate a differential boost in the accommodated condition on either the computations or concepts and applications measures. On problem-solving measures, however, students with disabilities profited more than students without disabilities when they were provided extended time, reading, or encoding accommodations.

When teachers plan accommodations for students, they are trying to predict what students will need in the testing situation to remove irrelevant barriers to performance. They are not predicting how it will affect test scores per se, but are trusted to make judgments as to whether or not the accommodation would remove irrelevant performance barriers yet not invalidate a test score. If, for example, an accommodation would likely invalidate test results (e.g., reading aloud tests of reading comprehension to students), they likely would not provide that accommodation. Fuchs and colleagues (2000) examined how teacher judgments of accommodations compared to a validation process they have developed. The researchers found poor correspondence between the number of accommodations that teachers recommended and the number of accommodations that their Dynamic Assessment of Test Accommodations (DATA) validation process awarded to students. In addition, in many cases, the students for whom teachers recommended accommodations often did not demonstrate a differential boost in performance, whereas students for whom the teachers did not recommend accommodations demonstrated greater accommodation boosts.

Fuchs and colleagues (2000) discussed the possibility that students who demonstrated differential boosts when accommodations were provided were likely students who possessed greater competence in the domain measured by the assessment(s). They hypothesized that in such cases, students were better able to take advantage of an accommodation and, as such, the accommodation removed construct-irrelevant variance because the disability was not intertwined with what the tests were measuring (i.e., mathematical competence). Fuchs and colleagues encouraged future researchers in this area to continue to explore (a) objective methods of validating accommodation decisions (i.e., look for differential accommodation boosts) and (b) additional demographic markers related to differential accommodation boosts for students with disabilities.

A question that comes up repeatedly when examining the Fuchs and colleagues (2000) study is: What is the impact of more than one accommodation on student performance? Fuchs and colleagues evaluated the impact of only one accommodation at a time, not a package of accommodations, on students' test scores. Given that previous research has documented that students rarely get just one accommodation, it is reasonable to question the generalizability of these researchers' findings. Perhaps students who did not demonstrate a differential accommodation boost need a combination, or package, of accommodations to remove their barriers to performance and thereby remove construct-irrelevant variance.

An investigation by Elliott, Kratochwill, and McKevitt (2001) focused on the use and effects of testing accommodations on the scores of students with disabilities on challenging mathematics and science performance assessment tasks. The major objectives of the investigation were to (a) document the testing accommodations educators actually use when assessing students with performance assessment tasks and (b) examine the effect accommodations have on test results. Both descriptive and experimental methods were used to analyze data. Individual cases of students with disabilities represent the strength and uniqueness of this research. The predominate research design in this investigation was an alternating treatments design (ATD). This data collection and analysis plan guided work with 100 fourth-graders, of whom 41 were students with disabilities. The results of the investigation indicated that slightly more than 75% of the testing accommodations packages that were suggested by students' IEP teams had a moderate to large effect (effect sizes of .50 to .81) on their test scores. It was also found that testing accommodations, to a lesser extent, had a positive effect on the test scores of students without disabilities. For a small percentage of students who were identified as exhibiting behavioral disorders, the effects of suggested accommodations were not positive.

Following the single-subject methodology of Elliott and colleagues, Schulte (2000) conducted a study that focused on the use and effect of testing accommodations on the scores of students with and without disabilities on alternate forms of a mathematics test typically used in statewide assessment programs. Her sample included 86 fourth-grade students including 43 students with disabilities and 43 students without disabilities. This study featured a 2 × 2 × 2 mixed design. All participants were tested under a treatment condition (i.e., accommodations during test) and a control condition (i.e., no accommodations during test). Testing conditions were randomized to combat the potential for order effects. Results indicated that both students with and those without disabilities, as groups, experienced a beneficial effect from testing accommodations. Although students with disabilities experienced a larger effect in the accommodated condition than students without disabilities, the difference between groups was not statistically significant. Students with disabilities experienced a small to medium effect (mean effect size .40), and students without disabilities experienced a minimal effect (mean effect size .25). Similar numbers of students with and without disabilities experienced either a beneficial effect, a detrimental effect, or a minimal to no effect in the accommodated testing condition. Not all students with disabilities unilaterally perceived the accommodated condition as better than the nonaccommodated condition, but most students without disabilities perceived no differences between accommodations or actually preferred the nonaccommodated condition.

Secondary analyses indicated that (a) students with disabilities who did not have math goals on their IEPs experienced a larger effect in the accommodated condition than did students with disabilities who did have math goals on their IEPs; (b) the accommodation package of extra time and read test/items to student did not have a differential impact for students with disabilities when compared to students without disabilities; (c) students receiving accommodation packages other than just extra time and read test/items to student experienced a statistically significant and differential impact of testing accommodations on math scores; and (d) students with disabilities profited more than students without disabilities on the multiple-choice items, but not on the constructed response items, as demonstrated by the interaction analyses and effect size statistics.

In summary, it seems clear that there is a need for rigorous empirical research in a variety of areas related to effects of testing accommodations on test scores. Specifically, testing accommodation research should include students with and without disabilities, testing both groups under accommodated and nonaccommodated conditions to investigate whether accommodations remove disability barriers or artificially raise performance (Trimble, 1998). Efforts also should be made to investigate the impact of individualized accommodations rather than focusing solely on "prepackaged" accommodations that may not be appropriate for every student. Furthermore, to completely understand the impact of testing accommodations on test scores, study designs need to examine individual effects via single-case methods as well as group effects. Of those reviewed, only the Fuchs and colleagues (2000), Elliott and colleagues (2001), and Schulte (2000) studies examined the effect(s) of accommodations on individual students. The vast majority of researchers have used group designs and in so doing may have lost information about individual effects. For research-based practices to emerge from this work, increased use of designs that also examine individuals' performances are needed.

Selecting and Using Testing Accommodations

By now, you should have a good understanding of what testing accommodations are and how they should function to improve the validity of a student's test score. In addition, you should be aware that testing accommodations are sanctioned by federal and state policies, and that IEP team members are responsible for selecting and implementing them for eligible students. But you can legitimately ask: "How do you go about selecting specific testing accommodations for specific students with specific disabilities and well-defined instructional plans?" The key to selecting and implementing testing accommodations for an individual student lies in the classroom(s) where that student is taught each day. That is, the instructional accommodations that teachers frequently use to facilitate the teaching–learning interactions for a student are prime candidates as accommodations when that same student is participating in a statewide or districtwide test. This premise is reasonable, particularly when there is good alignment between what is taught in the classroom and what is on the test. This does not mean, however, that all accommodations used to support a student during instruction will result in valid testing accommodations. More will be said about selecting and implementing testing accommodations later in this chapter via illustrations from two of our case studies.

The central role of testing accommodations is to improve the validity of the inference one makes from a test score of a student who has a disability when that disability involves abilities other than those being directly measured. In other words, testing accommodations are intended to improve the measurement accuracy of the test for students with disabilities and thus make their scores in math, reading, or other subject matter areas comparable to those of all other students. In many cases, for students with disabilities who are appropriately accommodated, researchers have found that their test scores increased when compared to their performances on a similar test when they did not receive accommodations (Elliott et al., 2000). This finding needs to be replicated by other researchers, but it makes sense logically given the intended role and function of testing accommodations.

Before concluding this section, it should be noted that the use of testing accommodations for students with disabilities has consequences for a number of parties. First the appropriate use of testing accommodations is clearly increasing the number of students with disabilities who participate in state- and districtwide assessments and probably is improving the resulting test scores for a significant proportion of these students as well. Second, test publishers are now more concerned about and interested in including students with disabilities in the standardization of their tests. Historically, standardization samples might be comprised of 4% or 5% of students with disabilities. Today the trend is more likely that standardization samples are comprised of 10% to 12% of students with disabilities. Third, because more students with disabilities are participating in assessments, educational stakeholders will be able to get more information on how these students are performing in school. Historically, a student's IEP was a process-focused document. Today, with the requirement that all students participate in large-scale assessments, the IEP is becoming an outcomes-oriented document. Fourth and finally, there is the potential for misuse. That is, some parents and students may be motivated to seek the inappropriate use of testing accommodations as a means to increase the likelihood of passing high-stakes examinations. Although no solid data exist that parents and students are doing this, educators should be aware of the possibility that students will be referred for special education services solely because of concerns about their test-taking skills and low achievement levels.

Before leaving our examination of testing accommodations, think about the wisdom of the "Do's and Don'ts in Testing Accommodations" offered by Thurlow, Elliott, and Ysseldyke (1998, pp. 61–62):

- *Don't* introduce a new accommodation for the first time for an assessment.
- *Don't* base the decision about what accommodations a student will use on the student's disability category.
- *Don't* start from the district or state list of approved accommodations when considering what accommodations a student will use in an upcoming test.
- *Do* systematically use accommodations during instruction and carry these into the assessment process.
- *Do* base the decision about accommodations, both for instruction and for assessment, on the needs of the student.
- *Do* consult the district or state list of approved accommodations after determining what accommodations the student needs. Then, reevaluate the importance of the accommodations that are not allowed. If they are important for the student, request their approval from the district or state.

As you work with students to provide testing accommodations, revisit this list and try to add to it. Now it is time to examine another assessment tactic, alternate assessments designed to facilitate the participation of students with some of the most severe disabilities.

Alternate Assessments: The Ultimate Accommodation

For many students with severe disabilities, changes beyond test administration procedures or format changes are needed to ensure that assessment results are meaningful. Thus, the content of the assessment also must be modified to provide for a valid measure of what these students are learning. This approach has led to the development of *alternate assessments* for approximately 15% to 20% of students with disabilities who are functioning at developmental and instructional levels significantly below those assessed by tests such as the Iowa Tests of Basic Skills, the Stanford 9, or the TerraNova.

Definition and Purpose

An alternate assessment is an assessment used in place of a state's or school district's regular achievement test (Ysseldyke & Olsen, 1999). Procedures for conducting an alternate assessment were still evolving in most other states as this book was being written even though IDEA required implementation of these assessments by July 1, 2000. Generally, an alternate assessment is understood to mean an assessment designed for those students with disabilities who are unable to participate in general large-scale assessments even when accommodations are provided. Thus, an alternate assessment is another tactic offered by IDEA and supported by state regulations that facilitates the inclusion of students with the most significant disabilities in assessment programs.

According to Heumann and Warlick (2000), on behalf of the U.S. Department of Education, "alternate assessments need to be aligned with the general curriculum standards for all students and should not be assumed appropriate only for those students with significant cognitive impairments. The need for alternate assessments depends on the individual needs of the child, not the category of the child's disability" (p. 8).

The number of alternate assessments is a state decision. As in many state- and districtwide assessment programs, the assessment may consist of multiple components or batteries. Title I requires that, at a minimum, reading/language arts and math must be assessed. Again according to Heumann and Warlick (2000),

> the alternate assessment should at a minimum assess the broad content areas such as communication, mathematics, social studies, science, etc. . . . The alternate assessment may assess additional content, including functional skills . . . . Functional skills can also be aligned to State standards as real work indicators of progress toward those standards. (p. 9)

The development and use of alternate assessments are evolving differently across the country, as attested to by the April 2000 NCEO cyber-survey on alternate assessment. The survey data indicated that states are aligning the con-

For More Information on
Alternate Assessment Practices,
see:
http://www.coled.umn.edu/NCEO

tent standards assessed by their alternate assessment to varying degrees with those assessed for general education students. For example, 6 states reported that "the standards are/will be identical to those applied to general education," 16 states reported that "the standards are/will be a subset of those applied to general education," and 8 states reported that "the standards have been/will be independently developed for students needing alternate assessments." However, 15 states indicated that they were "uncertain at this time" about the alignment of their alternate assessment with state content standards. Approximately 50% of the states reported that they were addressing issues concerning eligibility guidelines, assessment instruments, scoring, and proficiency levels for interpreting results of alternate assessments. However, most states reported that they had not addressed matters concerning the inclusion of scores in high-stakes systems or training for implementation of their alternate assessment system. Finally, when asked what "assessment approaches have been considered to date," states most frequently responded: direct observation, personal interview, behavioral rating scales, analysis and review of progress, or student portfolios.

Regarding the assessment method used, it is clear that teachers of students with significant disabilities will need to play an important role in the ongoing collection and interpretation of evidence that is indicative of the academic standards in their particular state. This activity has implications for how teachers write IEPs and the focus of their instruction of students with significant disabilities.

As indicated by the NCEO survey and our experience in numerous states, it appears that a majority of states are borrowing heavily from technology used in the development of behavior rating scales or performance and portfolio assessment. These technologies are based on teacher observations and the collection of student work samples. These methods, if used appropriately, have the potential to offer statistically sound results. More will be said about these assessment technologies when we examine research on alternate assessments.

A final trend that we are observing in alternate assessments across the country is that more students are taking alternate assessment in the area of reading than in any other area. This is due to the fact that fewer appropriate testing accommodations exist for students with serious reading disabilities. In addition, many educators may be incorrectly assuming that 8th- or 10th-grade large-scale tests require 8th- or 10th-grade reading skills to test successfully. This conclusion is wrong. Such tests have a range of readability levels that generally span two or three grade levels lower than the targeted grade.

Research on Alternate Assessment

As of 2000, very little research had been done under the name of alternate assessment. A review of the literature identified a few technical reports from research centers such as NCEO that describe alternate assessment practices in

Maryland and Kentucky, or the Mid-South Regional Resource Center that provide descriptions of alternate assessments in Delaware, Idaho, Indiana, Michigan, Missouri, North Carolina, and Tennessee, in addition to those on Kentucky and Maryland (Warlick & Olsen, 1999). Do not, however, conclude that there is not a research base for alternate assessments. In fact, the conceptual and measurement foundations for alternate assessment are well developed and are based on years of research in education and psychology covering performance assessment, behavioral assessment, developmental assessment, structured observations, and clinical assessment. Although these assessment methods differ somewhat, they all (a) are based on some direct or indirect observation of students, (b) are criterion or domain referenced in nature, and (c) require some summary judgments about the synthesis of data and the meaning of the scores or results. This latter quality, the use of judgments by knowledgeable assessors, is the empirical foundation for alternate assessment in states such as Indiana, Idaho, Pennsylvania, and Wisconsin. Therefore, a brief review of the research literature follows on the accuracy of teachers' judgments of students' academic functioning.

Hoge and Coladarci (1989) reviewed research on teacher-based judgments of academic achievement, consisting of 16 studies examining the relationships between teachers' judgments of student achievement and students' actual performances on an independent criterion of achievement. They concluded that "the results revealed high levels of validity for the teacher-judgment measures" (p. 297). Studies differed according to how the accuracy of teachers' judgments was assessed. The majority of the studies reported judgment/criterion correlations, and a few reported performance/judgment agreement data. The judgment/criterion correlations of the studies reviewed by Hoge and Coladarci ranged from .28 to .92. "The median correlation (.66) suggests a moderate to strong correspondence between teacher judgments and student achievement" (Hoge & Coladarci, 1989, p. 303). Hoge and Coladarci also compared the judgment/criterion correlations among the different methodological dimensions used. Indirect measures had a median correlation of .62, and direct measures had a median correlation of .69. On the dimension of judgment specificity, studies using rating scales had a median judgment/criterion correlation of .61. This was somewhat lower, although generally consistent with the correlations in studies using ranks (.76), grade equivalents (.70), number correct (.67), and item judgments (.70). Peer-referenced versus norm-referenced judgments did not seem to affect the judgment/criterion correlations. The peer-referenced median judgment/criterion correlation was .68, and the norm-referenced judgment/criterion correlation was .64.

A study by Gresham, Reschly, and Carey (1987) examined the accuracy of teachers in judging academic performance and in classifying students as having learning disabilities or being nondisabled. This study is relevant because of its examination of the accuracy of teachers' judgments. In the Gresham and colleagues study, the teachers' classifications were compared to the students' standardized test results. This study consisted of 100 children with learning disabilities and 100 children without disabilities. All of the students were given the Wechsler Intelligence Scale for Children-Revised (WISC-R) and the Peabody Individual Achievement Test (PIAT). Teachers were asked to fill out the Teacher Rating of Academic Performance (TRAP), a 5-item scale focusing on reading and math performance. The researchers reported that teachers' judgments of academic achievement were accurate in identifying students as having learning disabilities or as nondisabled. Furthermore, teachers' ratings on the TRAP identified children with learning disabilities somewhat more accurately than the WISC-R and the PIAT combined, 96% versus 91%. The opposite was true for the identification of students without disabilities; the WISC-R and the PIAT were slightly more accurate, 88% versus 86%. The researchers concluded that general classroom teachers are accurate "tests" of student academic achievement and could be used as one of the criteria by which psychoeducational tests are validated (Gresham et al., 1987).

In summary, information collected through alternate assessments is likely to be different from that collected for most students who take tests such as the Iowa Tests of Basic Skills, Stanford, or TerraNova, but if it is well aligned with the same academic standards, it still can serve as an index of student progress toward gaining skills that are held essential for all students in a given state.

Alternate Assessment Practices and Issues

Two states, Kentucky and Maryland, have been operating alternate assessments with some success for several years as part of a high-stakes state assessment system. These state assessments both emphasize performance assessments of academic and functional skills and require the use of portfolios that are scored by teams of raters using proficiency rubrics (Ysseldyke et al., 1996). Several articles have recently been written about how the Alternate Portfolio functions as part of the Kentucky Instructional Results Information System (see, for example, Kleinert, Haig, Kearns, & Kennedy, 2000; Turner, Baldwin, Kleinert, & Kearns, 2000). Let's take a closer look at the Kentucky Alternate Portfolio as an example of how alternate assessments are working.

With the passage of the Kentucky Education Reform Act in 1990, Kentucky became the first state to require full inclusion of special education students in large-scale assessments. The Alternate Portfolio is part of the Kentucky Instructional Results Information System and is designed as an option for students who have significant disabilities and are not working toward a regular diploma. This portfolio is designed to be generally aligned with the state's academic standards. To accomplish this developmentally downward extension of the state's standards, educators and parents identified "related critical functions." For example, for the academic standard "Students construct meaning through print for a variety of purposes through reading," the related critical function that students taking an alternate assessment were responsible for providing evidence of is "reads environmental pictorial print." Another example involves the standard of "Students use appropriate and relevant scientific skills to solve problems in real-life situations" and the related critical function of "Problem solves in new or novel situations."

Students in Kentucky are assessed when they are in 4th, 8th, and 12th grades. Thus, students with significant disabilities participate in an alternate assessment when they are at the same age-points as students in these grades. Features of the Alternate Portfolio include the following:

- The Alternate Portfolio includes evidence of how the student communicates, the student's daily and weekly schedule, and academic work sample entries that are aligned, downward extensions of the state's academic standards.

- The Alternate Portfolio is scored by two educators other than the child's teacher, using a scoring rubric.
- The scoring rubric is designed to describe a student's proficiency level (i.e., Novice, Apprentice, Proficient, or Distinguished) for each of six areas of functioning: (1) performance on targeted skills and participation in portfolio generation process; (2) use of natural environmental supports, including peers; (3) variety of settings in which performance occurs; (4) interactions with others; (5) use of multiple instructional types and contexts; and (6) reconciliation of key domain areas and concepts in state standards.
- Students' scores on the Alternate Portfolio carry equivalent weight to those of students in the regular assessment and are used collectively to calculate an overall school accountability index. School scores are used as the basis for rewards and recognition.

Kentucky's Alternate Portfolio is also designed to measure both student learning and the quality of instructional supports provided by the school. In many respects, the Alternate Portfolio is similar to the regular assessment that a majority of students take, because it also includes a portfolio component. The work in both portfolios, the one in the regular assessment and the Alternate one, are scored as Novice, Apprentice, Proficient, or Distinguished. The Alternate Portfolio is unique, however, because it reflects different performance areas than the regular portfolio.

At this point, you should have a good picture of what an alternate assessment could involve and what its role is in an educational accountability system that includes all learners. Let's now take a step back and examine the issues that impact the implementation and meaningful use of alternate assessments. These issues include (1) alignment with learning standards, (2) scoring of evidence, (3) time and timing, (4) parent involvement, (5) reliability and validity of results, (6) out-of-level testing, (7) information storage, and (8) reporting of results.

1. *Alignment with standards.* In general, all of the states we have explicitly mentioned are concerned that their alternate assessments are reasonably well aligned with their academic or learning standards. For example, Wisconsin and Idaho specifically emphasized IEP objectives as part of the alignment process. All three states also have reasonably clear criteria for making decisions about participation in an alternate assessment, but because the consequences for the assessment system in Kentucky are perceived to be higher (i.e., you cannot earn a diploma if you are in the alternate assessment), a lower percentage of students appear to participate in the Kentucky alternate assessment than in states such as Wisconsin or Idaho.
2. *Scoring of Evidence.* The scoring and reporting of the results of the alternate assessments clearly differs across states. In states such as Indiana, Idaho, and Wisconsin, the scoring of results is entrusted to IEP team members, whereas in Kentucky scoring is done by trained raters off site. Thus, in states where teachers score their own students who are

taking the alternate assessments, teachers appear to have more powerful and perhaps quicker feedback about students than teachers in places like Kentucky and Maryland, where alternate assessments are scored by educators who do not know the students. These latter states' scoring, however, may be more reliable, given that a teacher bias factor is reduced, if not removed, from the scoring.

No universal standard or consensus scoring system exists for evaluating the results of alternate assessments; yet we have observed that the rating scales and scoring rubrics used in several states have four aspects in common. First, scoring almost always focuses on the frequency with which a student performs a skill or task. The more frequent a student exhibits a skill, the more well developed the skill is thought to be. Second, a majority of scoring systems consider the amount of support a student needs to carry out a task or enact a skill. This is a common feature of instruction for students with significant disabilities. The smaller the amount of support needed, the more well developed the desired skill is thought to be. Third, scoring systems tend to value students' being able to exhibit a skill across multiple settings and with different people. This is referred to as the *generalizability* of the skill. The more generalizable the skill, the more well developed it is thought to be. Fourth and finally, evaluators of students' skills value the quality or accuracy with which a skill is exhibited. The more accurate or better the quality, the more well developed the skill is considered to be. The scoring approach a state uses influences the time it takes to score an assessment and the reliability of such scores. In general, scoring rubrics with the characteristics we have described can be used reliably by educators after some training and with periodic monitoring for accuracy.

3. *Time and timing.* The issues of time and timing are challenges that are always a consideration in any assessment, but in alternate assessments in particular. The collection of recent, representative, and reliable learning evidence by teachers and others means that these assessments should be an integral part of instruction. Although some states don't formally require that a portfolio be assembled, functionally most educators are collecting information over several weeks' time that will serve as the basis for evaluative judgments about student learning. In addition to the amount of time needed to collect and score information, educators must be cognizant of when they must report the results so that they can be integrated along with the test results of students participating in the state's regular assessment. Thus, the timing of the assessment can make the task of conducting an alternate assessment challenging.

4. *Parent involvement.* Parents are clearly expected to be recipients of the results of alternate assessments in each state. In states such as Indiana, Idaho, and Wisconsin, where the IEP teams play the major role in designing the assessment and collecting evidence, parents can play a significant role throughout the assessment.

5. *Reliability and validity of results.* The reliability and validity of the results of alternate assessments is a concern discussed in documents published by virtually all states. Remember, as we discussed in Chapter 2, consistency is central to the concept of reliability. In the case of an alternate assessment, where results are based on the judgments of educators who review an array of evidence about a particular student's learning, aspects of reliability concern the consistency among the judgments of IEP team members, the consistency of judgments over time (say, 3 or 4 weeks), and the agreement between educators' judgments of performance and actual test scores of students. Kentucky is an example of a state that has been the most vigilant about ensuring the reliability of its scoring method. Kentucky's approach has resulted in students' portfolios being scored by trained scorers who do not know the students. This step probably reduces bias in scoring and thus increases the reliability of the resulting proficiency characterization. The validity of all alternate assessments that emphasize the use of a portfolio is heavily influenced by the representativeness of the work samples and behavior evidence considered. In theory, these alternate assessments may yield highly valid results if educators do a good job of collecting and scoring representative samples of student work.
6. *Out-of-level testing.* None of the states we have examined closely advocate the use of out-of-level testing as a method for conducting an alternate assessment. Out-of-level testing means assessing students in one grade level using versions of tests that were designed for students in other, usually lower, grade levels. According to Heumann and Warlick (2000), IDEA does not specifically prohibit out-of-level tests, although they indicate that such a practice may be problematic for several reasons. One reason is that out-of-level testing may not assess the same content standards at the same levels as are assessed in the grade-level assessment. Also some assessment experts argue that out-of-level testing produces scores that are not comparable to those from the regular assessment and thus should not be aggregated (Bielinski, Thurlow, Minnema, & Scott, 2000; Minnema, Thurlow, Bielinski, & Scott, 2000). In theory, out-of-level tests could be used in a portfolio as part of the evidence about a student's academic functioning. However, few data suggest that this is happening.
7. *Information Storage.* A practical issue with alternate assessment is the storage of the information collected for each student. For example, in Idaho educators are expected to keep the results of alternate assessments at the local school level with student's cumulative files for 5 years. A written record in the IEP folder is the only record storage issue facing Wisconsin educators, whereas Kentucky's system appears to require storage of portfolios for several years, but it is unclear exactly how long and in what form. Educators in Indiana are required to use electronic portfolios, so their storage needs are greatly minimized.
8. *Reporting results of alternate assessments.* Scoring and reporting of alternate assessments are highly related, but because reporting in many ways represents the "bottom line," we have left our examination of this

issue until last. An alternate assessment in most states requires educators to understand their state's content standards, students' IEP objectives, and statistically sound methods for collecting achievement data on individual students. For many IEP teams, alternate assessments will result in the use of an array of methods for collecting individualized information that is recent, representative, and reliable. If the focus and subject matter coverage of the information collected is different across students, it should not be aggregated or summed together like performances on a test that is comprised of a common set of items. Thus, unless students taking an alternate assessment are given the exact same items under similar conditions, aggregating and reporting their scores for comparative purposes is questionable. In some states, aggregration is occurring at the global descriptive level. For example in Indiana, Wisconsin, and Idaho, for purposes of accountability at the state level, a student who completes an alternate assessment is said to be functioning at a "Prerequisite Skills Level." This approach allows all students to be described as functioning at one of several proficiency levels for each content area (i.e., reading, language arts, mathematics) assessed.

It seems that some educators are concerned that students with disabilities will score lower on tests than many other students, and consequently will lower the overall average score earned by a school and district. To address this concern that students with disabilities may lower a school's or district's scores, test scores for students with disabilities should be reported together with scores for their peers without disabilities and then be *disaggregated,* or reported separately, from those of other students. At this time, it is not considered appropriate to report which students received accommodations and which did not (McDonnell et al., 1997; Phillips, 1994) because of the possibility of *flagging* a student as having a disability.

In a recent report titled *Where's Waldo? A Third Search for Students with Disabilities in State Accountability Reports,* Thurlow and her associates (Thurlow, Nelson, Teelucksingh, & Ysseldyke, 2000) noted that even with the pressure of the IDEA public reporting requirements, only 14 states included participation data and only 17 states included performance data for students with disabilities in statewide assessments. Reporting remains a serious technical and social issue for states. Other issues confronting educators and states conducting alternate assessments will most certainly emerge as states across the country digest the results of their newly developed systems. These existing issues are not minor. If not addressed and handled well, the results of such assessments will not be meaningful. Clearly, more information is needed about how the various alternate assessment systems, mandated by law, are functioning.

Guidelines for Testing Students with Disabilities: Putting Testing Accommodations and Alternate Assessments into Practice

Up to this point, we have attempted to provide a legal, technical, and conceptual foundation—with a few Do's and Don'ts sprinkled in—for understanding testing accommodations and alternate assessments. It is now time to look into some of the details of putting this new knowledge into practice.

Many of the details for guiding your use of testing accommodations can be found in a document published by your state department of education or state office of educational accountability. As a starting point for this examination of practical steps for including all students with disabilities in assessment programs, here are several key recommendations highlighted in most of the participation and testing guidelines we have read.

For More Information on

Your state's testing guidelines, see:

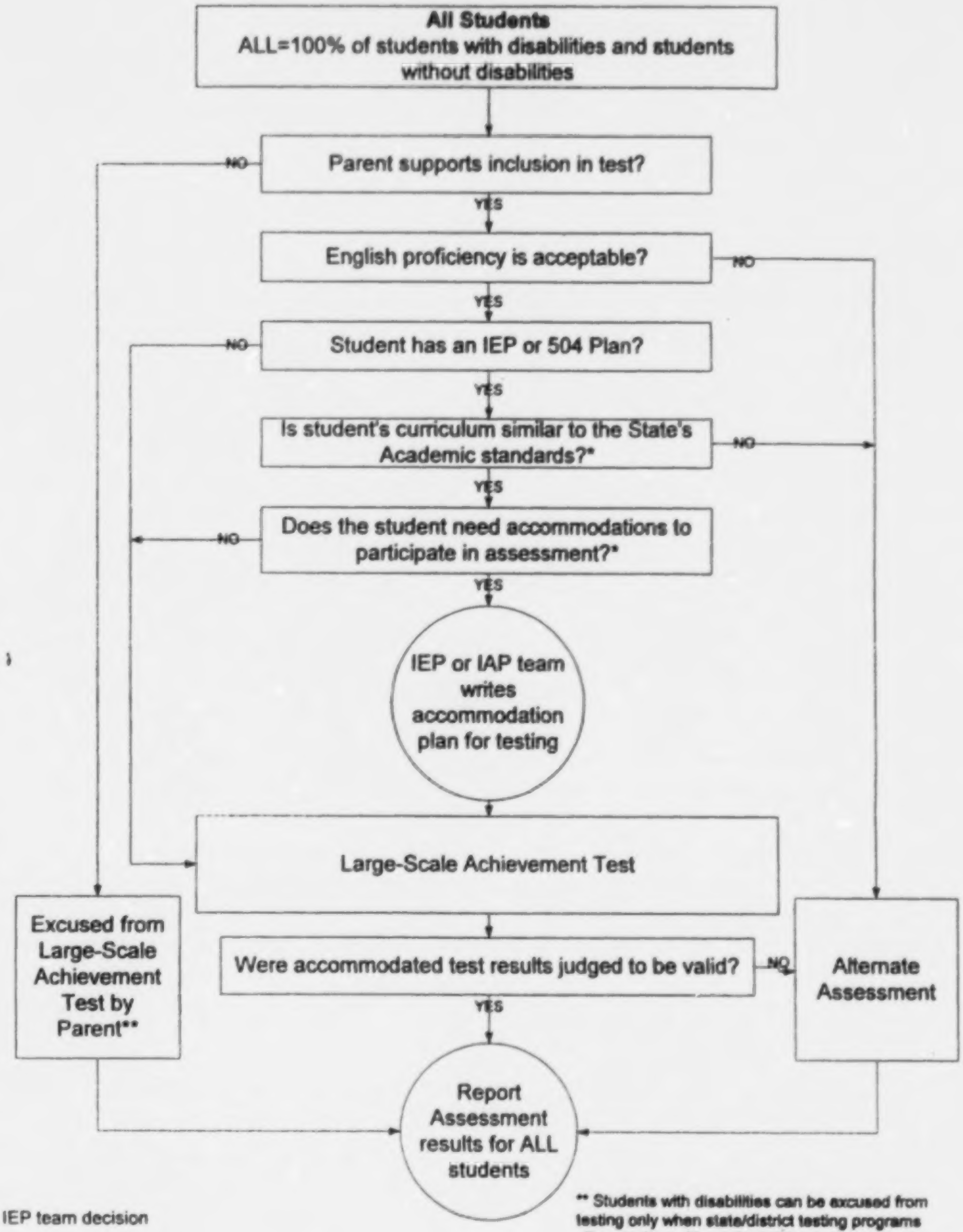
http://www.ccsso.org

- A student's IEP team, which includes the parent as an equal participant, addresses all questions regarding the participation of a student in state- and districtwide tests.
- State and federal special education law require that a student's IEP include statements of
  1. Whether or not the child will participate in the standardized test.
  2. Accommodations necessary to allow the child to participate in the test.
  3. If the child is not participating in the test, a statement of why the test is not appropriate and how the child will be assessed.
- To make these statements, the IEP team must know about the child's present level of educational performance and measurable annual goals, the general curriculum, the format and content of the state or district test, and the alignment between the curriculum and the academic content standards assessed by the state- or districtwide assessment system.
- Participation in the state (or district) test for students with disabilities is not an "all or nothing" decision. Instead, there are multiple options for enabling a student with a disability to participate. These options include participation in the test without accommodations, participation in the test with accommodations, and participation in alternate assessments.
- The IEP team decision regarding student participation in state assessment must be made on an individual basis. As a result, this decision is based upon a thorough review of child-specific data to assess the student's current educational performance relative to the academic performance standards for *all* students.
- This thorough review includes consideration of existing student records, including the most recent evaluation data, formal and informal evaluations conducted by team members, reports by parents and teachers, classroom work samples, independent educational evaluations, and any other information available to the IEP team.

- To make appropriate decisions regarding the student's need for accommodation and/or alternate assessment, the IEP team should consider the following: Begin with the assumption that the student will participate in the test and assess the need for accommodation and/or alternate assessment based on the student's present level of educational performance, IEP goals, and the content and format of the test. Consider the accommodations that the student receives in classroom assessments as possible accommodations for the test. Then select accommodations that *do not change the skills or content tested.* If the necessary accommodations would change the skills or content tested, the student's knowledge and skills should be assessed through alternate assessment. For example, an accommodation that included reading passages and/or items aloud to students would not be an acceptable accommodation if the purpose of the assessment is to measure reading skills. Thus, a student who would require this accommodation should participate in an alternate assessment to meaningfully assess his or her reading skills.
- Based on the thorough review of the student's current educational performance relative to the academic standards, the IEP team determines how a child with a disability will participate in the assessment system. For those students who are identified as needing accommodations on the standardized test, the IEP team must specify which accommodations are necessary for the child to participate in the assessment.
- The IEP team may determine that, even with accommodations, a child with a disability would be unable to demonstrate at least some of the knowledge and skills on the test. As a result of this decision, the student's performance will be assessed through alternate assessment.
- Test results are not the sole method for making educational decisions involving students with disabilities. Test results are only *part* of the information used to understand a learner and to monitor his or her educational progress.

The flowchart illustrated in Figure 4.3 was influenced by the state of Wisconsin's testing guidelines, but based on our review of other states' guidelines, it provides a good overall summary of the decision-making process surrounding students with disabilities in most states. Take a close look at this flowchart and try to use it to explain the assessment options for students with disabilities in your state.

It should be noted that several states—Alaska, Kansas, Oregon, and Rhode Island—make a general statement in their accommodation policies indicating any accommodation for any student is allowed. In addition, the state of Colorado's policy indicates that any student will be allowed to use any accommodation provided it has been in place for months prior to testing. The accommodation policies in both Maine and New York explicitly state that students who are ill or have acquired a temporary disability such as a broken wrist before the testing session may use accommodations without IEP documentation (Thurlow, House, Boys, Scott, & Ysseldyke, 2000).


All Students
ALL=100% of students with disabilities and students without disabilities
Parent supports inclusion in test?
NO
YES
English proficiency is acceptable?
NO
YES
Student has an IEP or 504 Plan?
NO
YES
Is student's curriculum similar to the State's Academic standards?*
NO
YES
Does the student need accommodations to participate in assessment?*
NO
YES
IEP or IAP team writes accommodation plan for testing
Large-Scale Achievement Test
Excused from Large-Scale Achievement Test by Parent**
Were accommodated test results judged to be valid?
NO
YES
Alternate Assessment
Report Assessment results for ALL students

* IEP team decision

** Students with disabilities can be excused from testing only when state/district testing programs allow student without disabilities to be excused.

FIGURE 4.3
Generic Flowchart of Questions and Decisions Concerning the Assessment of All Students in Statewide Assessment Systems

BEST COPY AVAILABLE

Case Applications

It is appropriate now to apply what we have discussed about testing accommodations and alternate assessments, so let's revisit the cases of Patrick, Tia, and Chris. Because Patrick lives in Florida and has not been identified as a student with a disability, he is ineligible for testing accommodations. In fact, as noted previously, Patrick would be denied testing accommodations in the vast majority of states, because testing accommodations are reserved for students with an indentified disability under IDEA or Section 504 or a student with limited English proficiency. Before examining the details of Tia's and Chris's cases, let's considered which type of assessment these two students should participate in given their respective grade levels, educational programs, and general competencies. For both Tia and Chris, the options include a test like the TerraNova, Iowa Tests of Basic Skills, or the Stanford 9, with or without accommodations; an alternate assessment; or some combination of a large-scale assessment and an alternate assessment. Educators who serve students with severe disabilities have reported that when they are making participation decisions it is helpful to focus on questions such as the following:

1. Is the student's curriculum very different from the district or state grade-level content standards? Yes or No?
2. Does the student demonstrate cognitive ability and adaptive behavior that prevent completion of the general education curriculum, even with program modifications and adaptations? Yes or No?
3. Are the student's management needs intensive, and does the student require a high degree of individualized attention and intervention from educators? Yes or No?
4. Does the student's current adaptive behavior require extensive direct instruction in multiple settings to accomplish the application and transfer of skills? Yes or No?
5. Is the student's inability to complete a course of study primarily due to his or her disability, rather than excessive or extended absences; language differences; or social, cultural, or environmental factors? Yes or No?
6. Is the student unable to apply or use academic skills at a minimal competency level in natural settings (e.g., home, community, work site)? Yes or No?
7. Does the student require intensive, frequent, and individualized community-based instruction to acquire, maintain or generalize skills and to demonstrate performance in settings such as prevocational and vocational settings? Yes or No?

These seven questions can serve as a participation decision checklist. When completed, they can serve as the basis for a justification to include or exclude a student from a large-scale assessment (Elliott & Braden, 2000). If four or more of the seven questions are answered "Yes," it seems unlikely that the results of a large-scale assessment would be meaningful even with appropriate testing accommodations. Today, all states with statewide assessments have written policies to facilitate participation decisions (see your state's policy or consult Thurlow, House, et al., 2000).

As you read about the cases of Tia and Chris, come back to our set of participation questions and see if you agree with the participation decisions made by these students' IEP teams. Remember, answering "Yes" to the majority of the seven questions serves only as a guideline for making participation decisions. In most cases, answering "Yes" to four or more of the points in the checklist would suggest that an IEP team believes that a student's cognitive capabilities are well below that of age-mates, that his or her curriculum is very different in content from what would be expected if it were reasonably well aligned with the state's content standards, and that the student needs extensive assistance to function at school and in other community settings. Thus, in effect, the content covered in each of the four subject matter areas of the test is highly likely to be very different from the subject matter in the student's daily curriculum. Consequently, to achieve a meaningful assessment of a student with a severe disability, the IEP team will have to utilize an assessment method other than a large-scale achievement test. For students for whom most of the responses to the checklist items are "No," it is highly likely that they can meaningfully participate in large-scale assessments with or without testing accommodations.

The Case of Tia

Recall that Tia is an eighth-grade student with a moderate learning disability, primarily difficulties in reading. She currently receives all her instruction in the regular classroom; however, her regular classroom teacher is supported by a consulting teacher who frequently helps to individualize some aspects of Tia's instructional tasks. Ms. DiPerna, Tia's regular teacher, stresses the use of authentic, performance tasks throughout instruction and assessment, particularly in mathematics and science. Ms. DiPerna also is quite knowledgeable of the state's content and performance standards in the areas of mathematics and science.

Tia is cooperative and motivated to do well. She works slower than most of her classmates because she reads slowly and has difficulty with composing written responses. Her IEP listed the following instructional accommodations: use of spelling aids to facilitate accuracy in spelling basic words, additional time to read and comprehend materials, read-along method to facilitate pace and comprehension of difficult text, and use of simple writing webs or diagrams to facilitate planning of written responses.

In preparation for the forthcoming IEP team meeting concerning Tia's participation in TerraNova, the state's large-scale assessment, Ms. DiPerna decided to try and discover which testing accommodations Tia would benefit from by administering several mathematics performance tasks used in previous years to evaluate all her students. She knew these tasks were challenging, requiring quite a bit of reading and spelling. However, based on her previous experience administering the state's test, she believed these tasks were a lot like many of the constructed response items on the mathematics and science tests. Therefore, she decided to administer the tasks to Tia with as many of her instructional accommodations in place as possible and then compare her

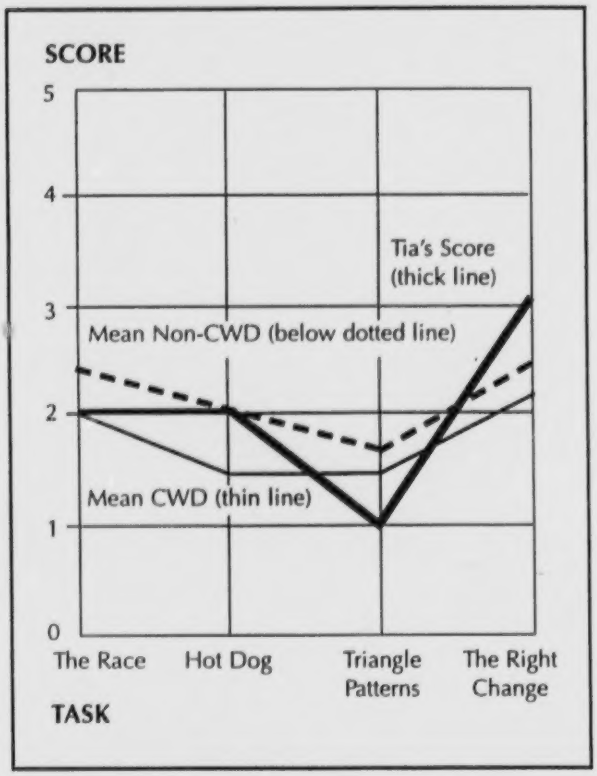


FIGURE 4.4
Tia's Scores on Ms. DiPerna's Mathematics Performance Tasks Compared to Other Students

Note. CWD means child with a disability. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 85. Copyright February 2000 Wisconsin Department of Public Instruction.

results to the mean scores of students without disabilities in her class. Tia was allowed extra time to read and respond to all the tasks, provided assistance with reading when she requested it, and allowed to use a dictionary and a spelling help sheet with many of her problem words written correctly. All the tasks that Ms. DiPerna used were scored using a rubric that had been posted in the room and which all her students, including Tia, understood. Specifically, the mathematics and science scoring rubrics were 0 = Not Scorable, 1 = Attempted Response, 2 = Minimal Response, 3 = Nearly Proficient Response, 4 = Proficient Response, and 5 = Advanced Response.

Figure 4.4 provides a summary of Tia's scores on the four mathematics performance tasks. In addition, Ms. DiPerna has included data from a previous class of students who also completed the same four performance tasks. This figure shows that Tia, with the use of accommodations that she was accustomed to during instruction, performed similar to the average of her peers without disabilities and above the average of other students with disabilities that Ms. DiPerna has taught over the past 2 years.

Once the tasks were completed, Ms. DiPerna asked Tia what she liked and didn't like about them. Ms. DiPerna also wanted to find out what Tia thought about the testing accommodations that she had used with her. The things Tia stated that she liked the most about the math tasks were that they asked interesting questions and that they were challenging. Regarding the aspects of the tasks she liked the least, Tia said that some of them needed more explaining and that the "triangle problem" was too complicated. She mentioned that there were parts of a few of the math tasks that she had never studied before. She suggested that the tasks might have been easier for her if they had provided more explanation of what students were expected to do and if she had more time to complete them. Armed with the data and knowledge from this practice testing experience, Ms. DiPerna listed the following possible testing accommodations for Tia:

- Provide extra testing time.
- Allow more frequent or extended rest breaks.
- Provide a distraction-free space or alternative location.
- Read and reread directions as needed.
- Clarify student questions about what to do by asking the student about what is written in the test booklet.
- Have the student reread directions to the teacher and restate them in his or her own words.
- Allow the special education teacher to administer the test.
- Read questions and content to the student.
- Give spelling assistance (use Spellmaster).
- Allow use of a calculator, manipulatives, and ruler.

With this information and the classroom testing experience with Tia, Ms. DiPerna felt ready for the forthcoming IEP meeting in which she knew testing accommodations were going to be discussed.

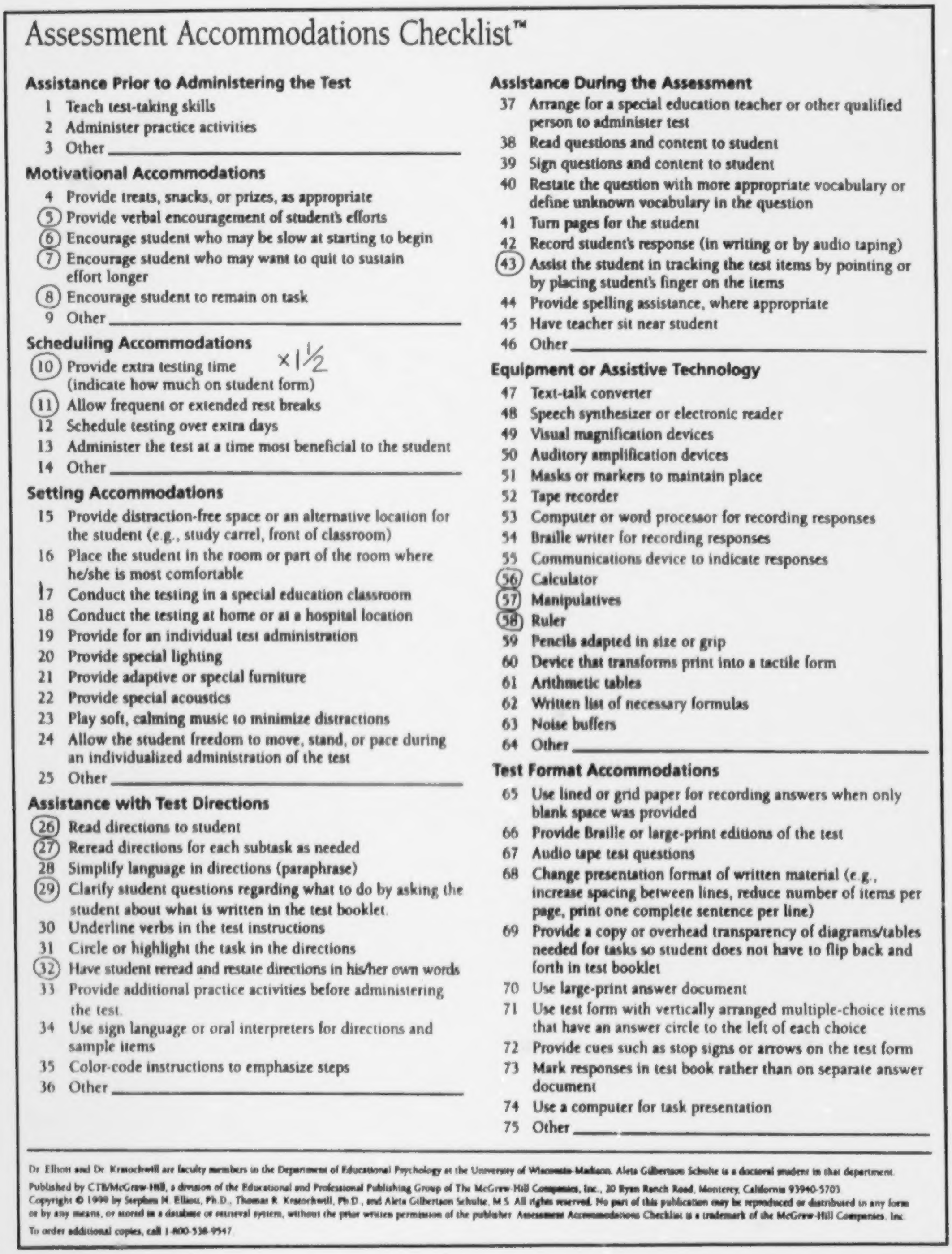
Tia's IEP team was required to meet to update her IEP with regard to participation in Wisconsin's statewide test and the possible need for testing accommodations. Tia's teacher in seventh grade was new to the school district and state, and subsequently had not felt comfortable at the end of the year making decisions about testing accommodations for Tia. Ms. DiPerna, Tia's mother, the school principal, and the school psychologist all met before the holiday break to discuss Tia's current educational functioning and her IEP goals, and specifically to make a decision about participation in the statewide test and the need for any testing accommodations.

To facilitate and focus participation at the meeting, the school psychologist, Mr. Roach, gave a brief overview of recent changes in federal and state law regarding the participation of all students in assessment programs and provided Tia's mother with a copy of a handout on testing accommodations. Tia's mother asked several questions about the state's test and why it was necessary for students with disabilities to be involved, given that they had already been tested quite a bit in the process of being identified with a disability. She also indicated that she was unaware of any state academic standards and requested a copy to review. After a rather lengthy discussion about the state's standards and the reasons for all students to participate in assessment programs such as the Wisconsin Student Assessment System (WSAS), the team addressed the issue of participation in the test. The IEP team answered each of the seven participation questions "No"; that is, the IEP team believed it was possible for Tia to participate meaningfully in the regular test, TerraNova.

The team then focused on identifying accommodations that would be needed to facilitate Tia's meaningful participation in the forthcoming test. At this point, Ms. ,DiPerna shared the results of her work with Tia. The team expressed interest in her findings but wondered about the applicability to the TerraNova of what she had learned about accommodating Tia from her classroom performance assessment tasks. Mr. Roach knew TerraNova well, given his testing expertise. Consequently, he was able to address questions about TerraNova and assured the team that although the items might differ in the content covered, many of the same skills needed to access Ms. DiPerna's performance tasks were similar to those needed to access the constructed response type items on the mathematics and science portions of TerraNova. At this point in the meeting, Mr. Roach reaffirmed that there was a consensus among the team that Tia should participate in the forthcoming statewide test and that she would need some accommodations to minimize the effect of her disability on the validity of the test results. Each team member voiced agreement with Mr. Roach, although it was clear that Mr. Kettler, the principal, had some reservations. Mr. Roach then introduced a copy of the Assessment Accommodations Checklist (see Figure 4.5) and noted that it could be used to help the team develop a testing accommodations plan and to communicate the plan with others who would be responsible for administering tests.

The team members agreed to try the Assessment Accommodations Checklist and came up with the following list of accommodations that they thought would be reasonable and would increase the validity of Tia's test scores on the Mathematics, Science, and Social Studies portions of TerraNova:

- Verbally encourage the student's effort.
- Provide 50% extra testing time.
- Allow more frequent or extended rest breaks.
- Provide a distraction-free space or alternative location for a small group.
- Read and reread directions as needed.
- Clarify student questions about what to do by asking the student about what is written in the test booklet.
- Have the student reread directions to the teacher and restate them in his or her own words.
- Allow the special education teacher to administer the test.
- Read questions and content to the student.
- Assist the student to track the test items by pointing or placing the student's finger on the items.



Assessment Accommodations Checklist™

Assistance Prior to Administering the Test

1 Teach test-taking skills
2 Administer practice activities
3 Other ____________

Motivational Accommodations

4 Provide treats, snacks, or prizes, as appropriate
[✓ 5] Provide verbal encouragement of student's efforts
[✓ 6] Encourage student who may be slow at starting to begin
[✓ 7] Encourage student who may want to quit to sustain effort longer
[✓ 8] Encourage student to remain on task
9 Other ____________

Scheduling Accommodations

[✓ 10] Provide extra testing time ×1½ (indicate how much on student form)
[✓ 11] Allow frequent or extended rest breaks
12 Schedule testing over extra days
13 Administer the test at a time most beneficial to the student
14 Other ____________

Setting Accommodations

15 Provide distraction-free space or an alternative location for the student (e.g., study carrel, front of classroom)
16 Place the student in the room or part of the room where he/she is most comfortable
17 Conduct the testing in a special education classroom
18 Conduct the testing at home or at a hospital location
19 Provide for an individual test administration
20 Provide special lighting
21 Provide adaptive or special furniture
22 Provide special acoustics
23 Play soft, calming music to minimize distractions
24 Allow the student freedom to move, stand, or pace during an individualized administration of the test
25 Other ____________

Assistance with Test Directions

[✓ 26] Read directions to student
[✓ 27] Reread directions for each subtask as needed
28 Simplify language in directions (paraphrase)
[✓ 29] Clarify student questions regarding what to do by asking the student about what is written in the test booklet
30 Underline verbs in the test instructions
31 Circle or highlight the task in the directions
[✓ 32] Have student reread and restate directions in his/her own words
33 Provide additional practice activities before administering the test
34 Use sign language or oral interpreters for directions and sample items
35 Color-code instructions to emphasize steps
36 Other ____________

Assistance During the Assessment

37 Arrange for a special education teacher or other qualified person to administer test
38 Read questions and content to student
39 Sign questions and content to student
40 Restate the question with more appropriate vocabulary or define unknown vocabulary in the question
41 Turn pages for the student
42 Record student's response (in writing or by audio taping)
[✓ 43] Assist the student in tracking the test items by pointing or by placing student's finger on the items
44 Provide spelling assistance, where appropriate
45 Have teacher sit near student
46 Other ____________

Equipment or Assistive Technology

47 Text-talk converter
48 Speech synthesizer or electronic reader
49 Visual magnification devices
50 Auditory amplification devices
51 Masks or markers to maintain place
52 Tape recorder
53 Computer or word processor for recording responses
54 Braille writer for recording responses
55 Communications device to indicate responses
[✓ 56] Calculator
[✓ 57] Manipulatives
[✓ 58] Ruler
59 Pencils adapted in size or grip
60 Device that transforms print into a tactile form
61 Arithmetic tables
62 Written list of necessary formulas
63 Noise buffers
64 Other ____________

Test Format Accommodations

65 Use lined or grid paper for recording answers when only blank space was provided
66 Provide Braille or large-print editions of the test
67 Audio tape test questions
68 Change presentation format of written material (e.g., increase spacing between lines, reduce number of items per page, print one complete sentence per line)
69 Provide a copy or overhead transparency of diagrams/tables needed for tasks so student does not have to flip back and forth in test booklet
70 Use large-print answer document
71 Use test form with vertically arranged multiple-choice items that have an answer circle to the left of each choice
72 Provide cues such as stop signs or arrows on the test form
73 Mark responses in test book rather than on separate answer document
74 Use a computer for task presentation
75 Other ____________

Dr. Elliott and Dr. Kratochwill are faculty members in the Department of Educational Psychology at the University of Wisconsin-Madison. Aleta Gilbertson Schulte is a doctoral student in that department.

Published by CTB/McGraw-Hill, a division of the Educational and Professional Publishing Group of The McGraw-Hill Companies, Inc., 20 Ryan Ranch Road, Monterey, California 93940-5703

Copyright © 1999 by Stephen N. Elliott, Ph.D., Thomas R. Kratochwill, Ph.D., and Aleta Gilbertson Schulte, M.S. All rights reserved. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher. Assessment Accommodations Checklist is a trademark of the McGraw-Hill Companies, Inc.

To order additional copies, call 1-800-538-9547.

FIGURE 4.5
Accommodations Selected from the Assessment Accommodations Checklist by Tia's IEP Team Members

From *Wisconsin Student Assessment System* (based on the Terra/Nova tests), 1997, Monterey, California: CTB/McGraw-Hill. Copyright 1997 by CTB/McGraw-Hill. Reproduced with permission of CTB/McGraw-Hill.

BEST COPY AVAILABLE

- Encourage the student to begin, remain on task, and sustain effort longer before quitting.
- Allow the use of a calculator, manipulatives, and ruler.

There was more disagreement about the accommodations Tia needed to meaningfully participate in the Reading and Language Arts test. They knew, of course, that the content of the items could *not* be read to Tia, but all the educators felt that it might be reasonable to read the possible answer foils on the multiple-choice items. In addition, there was some debate about the amount of time Tia would need to complete the test. Ultimately, the team endorsed the same list of accommodations for the Reading and Language Arts test with the exception of the accommodation of "reading questions and content to the student."

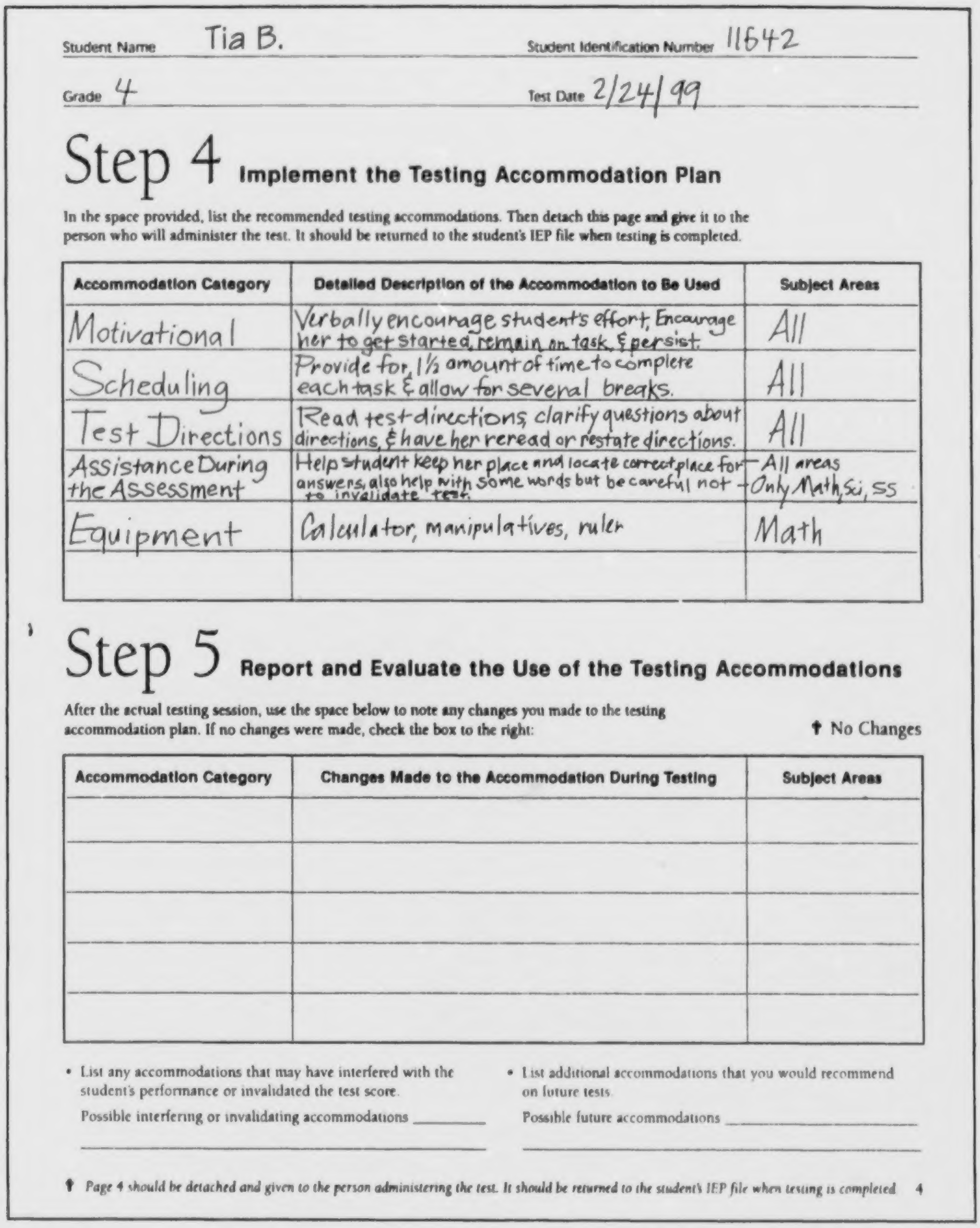
As a result of the IEP team meeting, a feasible testing accommodation plan was developed that should facilitate Tia's meaningful participation in the forthcoming statewide test. Implementation of the plan will require the attention of a test administrator who is responsible for only a few students and a testing setting where communication between Tia and the test administrator can occur without disrupting other test takers. A copy of the testing accommodation plan summarized from the Assessment Accommodations Checklist for a test administrator is shown in Figure 4.6. If these accommodations are carried out, it is the professional judgment of the IEP team that the resulting scores will be better indicators of Tia's abilities in mathematics, science, social studies, and reading and language arts. Thus, the accommodation plan is designed to increase the likelihood that Tia will actually take the test and that her scores will provide a valid indication of her abilities

The Case of Chris

Chris, as you recall, is an 11th-grade student with Down syndrome who attends high school in Idaho. He receives the majority of his instruction in a highly structured classroom with 10 other students, his teacher, Mrs. Davidson, and her teaching aide. Like many students with Down syndrome, Chris's interpersonal skills are immature but acceptable. He receives extensive instructional support and spends the majority of his school day working on functional communication and daily living skills. One of his favorite classes is "Employability Class," where he recently began a job helping to clean the lunchroom.

Chris's IEP team concluded that he should be given an alternate assessment due to the pervasive nature of his disability and the fact that his current educational curriculum was very different from the curriculum of a majority of his age-mates. Specifically, Chris's IEP team members answered six of the seven participation decision checklist questions "Yes" when reflecting on his work in reading, mathematics, science, and social studies, so they recommended that Chris not participate in the Iowa Tests of Basic Skills (ITBS), Idaho's large-scale assessment. In place of the ITBS, the IEP team decided to conduct an alternate assessment.

Mrs. Davidson volunteered to provide leadership in conducting the alternate assessment. Each of the other team members, Dr. Carroll (the school psychologist), Ms. Wayley (the principal), and Chris's mother agreed to help



Student Name Tia B.
Student Identification Number 11642
Grade 4
Test Date 2/24/99

Step 4 Implement the Testing Accommodation Plan

In the space provided, list the recommended testing accommodations. Then detach this page and give it to the person who will administer the test. It should be returned to the student's IEP file when testing is completed.

Accommodation Category	Detailed Description of the Accommodation to Be Used	Subject Areas
Motivational	Verbally encourage student's effort, Encourage her to get started, remain on task, & persist.	All
Scheduling	Provide for 1½ amount of time to complete each task & allow for several breaks.	All
Test Directions	Read test directions, clarify questions about directions, & have her reread or restate directions.	All
Assistance During the Assessment	Help student keep her place and locate correct place for answers, also help with some words but be careful not to invalidate test.	All areas Only Math, Sci, SS
Equipment	Calculator, manipulatives, ruler	Math

Step 5 Report and Evaluate the Use of the Testing Accommodations

After the actual testing session, use the space below to note any changes you made to the testing accommodation plan. If no changes were made, check the box to the right: † No Changes

Accommodation Category	Changes Made to the Accommodation During Testing	Subject Areas

- List any accommodations that may have interfered with the student's performance or invalidated the test score.

  Possible interfering or invalidating accommodations ____________

- List additional accommodations that you would recommend on future tests.

  Possible future accommodations ____________

† *Page 4 should be detached and given to the person administering the test. It should be returned to the student's IEP file when testing is completed* 4

FIGURE 4.6
Summary of Tia's Accommodation Plan as Written on the Assessment Accommodations Checklist Form

From *Wisconsin Student Assessment System* (based on the Terra/Nova tests), 1997, Monterey, California: CTB/McGraw-Hill. Copyright 1997 by CTB/McGraw-Hill. Reproduced with permission of CTB/McGraw-Hill.

BEST COPY AVAILABLE

Mrs. Davidson, but they felt somewhat at a loss as to what test to use to assess Chris's skills in mathematics, reading, science, and social studies and how to accommodate him during the test. Mrs. Davidson explained to the team members that an alternate assessment in Idaho involves the use of the Idaho Alternate Assessment (IAA). The IAA is a behavior rating scale that focuses on the evaluation of a student's IEP objectives that are well aligned with the state's alternate knowledge and skills standards. Mrs. Davidson indicated that the team's main task would be to examine the rather substantial collection of classroom work samples that Chris had produced in mathematics and reading and to review the weekly notes that she and her aide had written over the past 3 months. Most of these progress notes concerned learning objectives on Chris's IEP and focused on communication skills, social skills, and employability skills. In addition to these notes, Ms. Willems, the classroom aide, had videotaped Chris during three instructional sessions when he was working on reading and writing.

With Mrs. Davidson's leadership, the IEP members agreed to review and rate the materials that had been collected and served as evidence of Chris's current knowledge and skills. However, Mr. Carroll, the school psychologist, questioned whether the evidence was enough. He commented that it seemed as if the evidence was recent and representative of what Chris had been doing in mathematics and reading or language arts, but he didn't see any evidence of work in science or social studies. Chris's mother disagreed mildly; she felt that the objectives on his IEP concerning social skills and employability skills were basic social studies skills. This point provoked quite a bit of discussion among the team members and generated a number of questions that nobody could answer with confidence. For example, if Chris's IEP didn't have any learning objectives concerning science, did the alternate assessment still have to document his achievements in science? How far can one work downward developmentally from the state's content and performance standards and still be assessing skills in mathematics or reading? How does one reliably score Chris's performance, and how are scores reported?

Mr. Carroll admitted that he was getting confused and a little uncomfortable doing an alternate assessment. He commented that there is error in any measurement; that is, all assessments have some error. But it seems as if an alternate assessment can be full of error, and the resulting scores might be meaningless given that every student could have a different assessment. Mrs. Davidson responded politely but firmly to Mr. Carroll's comments about error and the potential for meaningless scores. She indicated that she had been a teacher for 18 years and had been responsible for evaluating the performances of hundreds of students. She went on to indicate that there is strong evidence that teachers can be excellent judges of students' work. Thus, teachers' judgments can be reliable and valid. Mrs. Davidson reminded the team that she was aware that several researchers have published work on the validity of teachers' judgments of student achievement. In addition, she indicated that research had also demonstrated that the use of scoring rubrics were tools that could enhance the reliability of teachers' evaluation of students' work in language arts and mathematics.

The IEP team members nodded their agreement with Mrs. Davidson. Therefore, she continued on and suggested that over the course of the next 2 weeks she and her aide would organize the evidence they had collected about Chris's learning and academic progress in the areas of mathematics, reading, and social studies. They would also review what if anything Chris had done in the area of science. Thus, another meeting was scheduled for about a month before the ITBS when the entire team could get together to do ratings of Chris's work via the IAA and then provide a summary report of their results for purposes of statewide and local accountability.

At the meeting the next month, Ms. Wayley, the principal, started the session with a review of the state's policy on alternate assessment and an overview of the state's academic content standards. With this information as background, the team agreed on three main points at the outset of the meeting:

1. For the purposes of statewide accountability reporting, Chris was functioning at the Prerequisite Skills level in mathematics, reading/language arts, social studies, and science even though he was not really doing any classwork on any of the science standards.
2. Substantial evidence (e.g., classroom work samples, teacher's progress notes, videotapes of communication skills, parents' observations) concerning Chris's academic functioning had already been collected that was deemed to be recent and representative of his work.
3. The main challenge the team faced was reliably rating and interpreting the evidence for instructional use.

Mrs. Davidson presented most of the evidence about Chris's functioning in a portfolio notebook, which made it easy to review and to see the date when the work was completed. She also had organized the information by subject matter and provided partially completed IAA forms that indicated how Chris's IEP objectives aligned or didn't align well with state academic standards for 11th-graders. As a means of interpreting the evidence and communicating with others about the quality of Chris's work, Mrs. Davidson used the IAA's scoring rubric that emphasized three dimensions: the frequency or quality with which a skill is exhibited, the range of settings in which a skill is exhibited, and the amount of support a student needs to exhibit the skill.

Dr. Carroll indicated that he was impressed with the rubric and was now able to more fully appreciate the points about reliable and valid scores that Mrs. Davidson had made at the earlier meeting where they discussed alternate assessment. Mrs. Davidson shared her relief to hear that her colleague liked what had been developed by the state and then turned to Chris's mother to see how she was reacting. She, too, was positive about the IAA rating scales because she thought they provided a meaningful measurement of what her son was learning. She understood that Chris was functioning at the Prerequisite Skills level in mathematics, reading, social studies, and science when compared to other students without disabilities. But she also recognized that this did not mean that Chris was not learning. There was substantial classroom-based evidence from Mrs. Davidson's classroom to prove that Chris was making progress. She concluded by saying that it seemed that more parents would

Case Applications and Good Assessment

Patrick

■ Patrick's state uses a commercially produced test that you learned about in the previous chapter. Suffice it to say that the test has been developed to yield reliable and valid scores. However, the fact that Patrick is a poor reader and is not eligible to receive any testing accommodations in Florida suggests that the validity of his scores on tests of mathematics, science, and social studies may not be highly valid indicators of his true skills in these subject matter areas. As you recognize, significant reading difficulties can influence any student's performance on a test on which reading is needed to access and use information.

Tia

■ Tia's state also uses a commercially produced test that is well aligned with state academic standards. This test was discussed in more detail in the previous chapter, but it, too, is well developed, and there is significant evidence that it generally yields reliable and valid test scores. Tia's reading disability, if appropriately accommodated, should not have a negative effect on the validity of her test scores in mathematics, science, and social studies. Her performance on the reading test, however, cannot be accommodated because reading is the skill that is being measured. Influencing her reading by using reading accommodations would result in invalidating the reading test score.

Chris

■ Chris's state uses a highly regarded commercially produced test that has substantial reliability and validity evidence. However, Chris will not be taking this test, because his curriculum focuses on functional skills. He will be taking an alternate assessment. Reliability and validity are still important issues to consider when reporting Chris's performance on Idaho's alternate assessment. Consequently, the educators conducting the alternate assessment will be responsible for documenting that the results are reliable and valid. Clearly, educators must really understand these technical concepts of good tests if they are going to conduct alternate assessments.

want their child to have an alternate assessment because it actually allows you to see what your child can and cannot do.

Ms. Wayley, who had been monitoring the discussion and the time of day, announced that the team still should spend a little more time on scoring Chris's work using the IAA scales. She encouraged each member of the team independently to look at the evidence and to select a number or level within the scoring rubric that best characterized the work. Once each member had done this for Chris's mathematics work, they shared their perceptions and discussed any disagreements. The consensus rating for Chris's mathematics work was characterized as "Developed," which resulted in a score of 3. A similar process was used to summarize his work in reading and social studies. In both of these subject matter areas, the team members came to a consensus characterization of "Developing," or a score of 2. With regard to science, there was no evidence to evaluate. Chris's IEP did not have any skills on it concerning science. The IEP team agreed that, according to the rubric, Chris's work was best characterized as "Nonexistent" or quantitatively a score of 0. As a result of this assessment, however, it was decided that when the IEP was reviewed at the end of the year there should be consideration of some basic skills in science.

With the completion of the scoring of Chris's evidence, Mr. Carroll encouraged the team to summarize its alternate assessment efforts in a brief report that could be placed in Chris's IEP file. Mrs. Davidson echoed this recommendation and suggested a simple report card type report, similar to what is sent home for students who take the ITBS when their tests have been scored. With the report completed, the team concluded they were finished with their alternate assessment of Chris. The assessment had provided them an opportunity to communicate about Chris's progress and to put it in the context of the state's academic standards. The team members felt that the alternate assessment was meaningful and provided valuable feedback to his parents and teachers. In addition, Chris and his assessment results were included in the state's accountability system, helping to provide a more complete picture of achievement of all students.

A Concluding Thought

Educators are now empowered and entrusted to include all students in the various assessment systems that have been implemented in their districts and states. To achieve this, new policies and practices have been advocated that involve testing accommodations and alternate assessments. In this chapter, we have examined in some detail the use of testing accommodations and alternate assessments as the two primary tactics available to facilitate the meaningful participation of students with special needs in assessments. Wise use of these assessment tactics rests upon an understanding of the concept of test score validity and an appreciation that good assessment is part of good instruction.

CHAPTER FIVE

Best Practices for Inclusive Assessment Programs and Educational Accountability

Today, perhaps more than ever, there is a strong interest in getting a clear and complete picture of how well students are learning and how well schools are functioning. Consequently assessing students—all students—is not only an important part of educational accountability, it also is the law. As you know, however, our schools educate a diverse group of students, among whom are students with disabilities. Meaningfully assessing the learning of all students with disabilities is a challenging task. Fortunately, the laws and regulatory procedures that guide the delivery of services for students with disabilities allow for the use of two assessment tactics, testing accommodations and alternate assessments, to facilitate the participation and meaningful assessment of these students. An understanding and intelligent use of these two assessment tactics within large-scale assessment programs has been the primary focus of this book. The application of these two tactics has been illustrated through the cases of Tia and Chris. It is now appropriate to take stock of what we have said and provide a summary of the state of the art of inclusive accountability efforts for students with disabilities.

Summary of Inclusive Assessment Practices

Decisions about the use of testing accommodations and alternate assessment need to be guided by common sense, state testing guidelines, and a sound understanding of test validity, because there is little published research to date on these tactics. To guide your use of testing accommodations, the following points have been stressed in this book:

- Decisions about testing accommodations for students receiving special education must be made by an IEP team and based on the *individual needs* of a student, not on the student's disability category.
- The testing accommodations to be used and those actually used on the student's IEP must be documented, and the IEP team's plan to accommodate a student must be communicated to his or her parents and the individual responsible for administering the test, if these people can't attend the IEP team meeting.
- Accommodations that a student currently receives during classroom instruction provide the starting point for selecting possible accommo-

dations that will facilitate test taking; using accommodations that a student has not experienced previously can actually create problems for the student.

- The purpose of a testing accommodation is to enhance the validity of the inference made from a student's test score; therefore, appropriate testing accommodations should impact access or enabling skills, not the skills or abilities targeted by the test.
- The list of known invalidating and nonstandard accommodations is actually quite short; it includes reading a reading test, using a calculator on a mathematics test that is designed as a measure of mental mathematics, using spelling aids on a test on which points are allocated for correct spelling, and using excessive paraphrasing of content that results in changing the meaning or level of difficulty of the material.
- IEP teams should meet to make testing accommodation plans several weeks prior to the actual test to ensure that testing personnel have time to coordinate accommodation plans for the entire group of students who need them.
- Testing accommodations must be reasonable and feasible—reasonable with respect to the number and type of accommodations that the student receives on a regular basis in his or her classroom and feasible in that the individual administering or managing the accommodations has the resources and skills to implement the accommodations accurately.
- If, after completing a test, you believe the accommodation(s) used invalidated the results, report it to the test coordinator and arrange for another administration without using the specific invalidating accommodation(s), or consider conducting an alternate assessment of the student.

When a student cannot meaningfully participate in a test, such as the reading portion or math portion of an achievement test, even with a comprehensive accommodation plan, an alternate assessment must be designed and administered to the student. To guide your use of alternate assessments, the following points were stressed in this book:

- IEP teams are responsible for making the decision about participation in an alternate assessment based on a series of issues, the utmost of which concerns the mismatch between the instructional level at which an individual student is working and the content and learning expectations characterized by the assessment. Decisions about participation should not be based on a student's disability category.
- IEP teams are responsible for conducting the alternate assessment, which at a minimum must involve a thorough and timely review of the student's achievement and progress on IEP objectives that are aligned with the academic standards framework for which all students in the state are held accountable. The focus of the alternate assessment can cover areas in addition to those embodied by the state standards.
- A variety of assessment methods, including observations, records reviews, work samples, performance tasks, and developmental or diag-

nostic tests, can be used to collect evidence to provide a basis for the assessment. The results of these assessments should be scored and/or summarized in writing, documented in the IEP, and stored for review by others—in particular, the student's parents and future teachers.

Inclusive accountability practices and federal law suggest that the assessment results for each student who participates in an alternate assessment be reported with the same frequency and level of detail and at the same time as the results of students participating in the regular assessment. Functionally, the primary reporting method of alternate assessment results for the public in most states will be information from schools about the number of students who took an alternate assessment and the fact that this is indicative of the students' functioning at a level commonly characterized as the Prerequisite Skills level. More detailed reports about students' achievements and progress should be provided to parents, but probably not aggregated in a summary, because the scores are not likely to be comparable given that the types of tasks and assessment methods used are often quite variable across students with severe disabilities.

Alternate assessment, like any other assessment, must be recent, reliable, and a representative sample of a students' skills and abilities. When these conditions are met and the content of the assessment is aligned with the state's content standards framework, the results can be interpreted with confidence.

Preparing All Students to Take Tests

Virtually all students will benefit from some test preparation practice and test-taking guidance. The goal of teaching is to increase learning rather than to increase test scores. Teachers, therefore, are reminded that students' attention and effort should be directed to learning the entire scope of the curriculum, not just the limited knowledge and skills measured by an achievement test. Some state testing guidelines actively discourage school staff from buying, developing, or promoting the use of extensive test practice materials that closely parallel the items or tasks on the state's test. Test preparation, these guidelines notwithstanding, is a frequent concern of many educators and parents. And given changes in requirements concerning the participation of all students in assessment programs and the emphasis on testing as a major aspect of promotion and graduation decisions, it is anticipated that test preparation efforts will increase. Therefore, we believe it is worthwhile to understand the role and ethics of test preparation for all students.

Many sound test preparation practices may appear to be common-sense activities; however, our experience with many educators suggests otherwise. Consider the test preparation strategies listed here, which some teachers reportedly use when administering tests such as the ITBS, Stanford 9, and TerraNova. As you read through the list, critically evaluate the strategies to determine which ones you believe are appropriate and which are not appropriate.

- Limit instruction during the month prior to the test to only those objectives that are thought to be on the test.
- During instruction use examples that are from last year's test.

- Give students an opportunity to practice taking the actual test items before they formally start the test.
- Teach students general test-taking skills (e.g., listen carefully to directions, read the entire question before answering) to improve their test performance.

To determine which of the four strategies are educationally and ethically sound, use two guiding principles:

1. The educational objectives, the content of instruction, and the content of the achievement test should be aligned or strongly related to each other.
2. The general purpose of an achievement test is to inform educators and students how well the students have learned what has been taught.

Thus, according to Airasian (1994), the important issue becomes just how strong the relationship should be among learning objectives (content standards), instruction, and the test. The National Council on Measurement in Education (NCME) task force (Canner et al., 1991) has provided a set of guidelines for what is appropriate and inappropriate test preparation. Its basic guideline states that all test preparation activities that lower the validity of interpretations made from test scores are inappropriate and should be avoided. The guidelines of the NCME task force indicate that the following test preparation activities are inappropriate or unethical:

- Focusing instruction only on task or item formats used on the test.
- Using examples during instruction that are identical to test items or tasks.
- Giving students practice taking the actual items on which they will be tested in the near future.

Ultimately, the issue of proper test preparation is one of validity. That is, the assessment of student achievement should provide a fair and representative indication of how well students have learned what they have been taught, and in order to do this, test questions must focus on knowledge and skills similar to those students were taught during instruction. Perhaps the most important word in the previous sentence is *similar*. There is an important ethical difference between teaching to the content standards a test measures and teaching the test itself! Teaching to the content standards that a test measures is a desirable practice; it involves teaching students the general knowledge and skills that they need to answer questions on the test and to succeed in future education and work settings. Teaching the test itself involves teaching students the answers to specific questions that will appear on the test. This is neither pedagogically appropriate nor ethical, because it can result in a distorted or invalid picture of what students have achieved.

Good test preparation should enable students to show what they have learned in classes over the past several years. Therefore, it is helpful for all students to understand that when taking a test they should:

- Be well rested and comfortable at the time of testing.

- Attend carefully to test directions and follow directions exactly.
- Ask questions when they are unsure of what to do.
- Find out how questions will be scored.
- Pace themselves so they do not spend so much time on some questions that they cannot get to other questions.
- Plan and organize essay questions before responding.
- Act in their own interest by attempting to answer all questions.
- When using a separate answer sheet, check often to make certain they are marking their responses accurately and in the correct place.

Besides these general test-taking guidelines, experts who study test taking, or what has become known as testwiseness suggest, some additional skills that provide students some strategies for answering test questions (Linn & Gronlund, 1995; Sarnacki, 1979). Most of these testwise skills relate to errors on the part of question writers who provide clues to correct answers. For example, when responding to multiple-choice questions, a testwise student knows that:

1. The answer option that is longest or most precisely stated is likely to be the correct one.
2. Answer choices that do not attach smoothly to the item stem are not likely to be correct.
3. The use of vague words such as *some, often,* or *similar* in one of the answer choices is likely to indicate the correct option.

In summary, many good test-taking skills can be mastered by virtually all students, but students need some practice to develop these skills and confidence in using them. Consider spending instructional time a couple weeks prior to an important test discussing and modeling good test-taking skills for all your students. Remember, however, test preparation should not raise test scores without also raising students' mastery of the general content being tested. Thus, test preparation and test-taking skills are designed to increase the validity of students' test scores, not necessarily to increase their scores.

Fair Testing Practices Require Efforts from Many People

Research suggests that teachers spend as much as a third of their time involved in some type of assessment. Teachers are continually making decisions about the most effective means of interacting with their students. These decisions are usually based on information they have gathered from observing their students' behavior and performances on learning tasks in the classroom and on standardized test results (Witt et al., 1998).

Many individuals have a vested interest in student learning and assessment information about such learning. Clearly, teachers, students, and parents should have great interest in the results of student assessments. School administrators and community leaders also voice keen interest in assessment results that document students' performances. No single assessment technique or testing procedure, however, can serve all these potential users of assessment results. Thus, the purpose of one's assessment must be clear, for it influences instruction and assessment activities and, consequently, the interpretation of any results.

Teachers have two main purposes for assessing students: to form specific decisions about a student or a group of students and to guide their own instructional planning and subsequent activities with students. Teachers use assessment results for specific decisions, including determining how well students have mastered what they have taught, diagnosing student strengths and weaknesses, grouping students for instruction, identifying students who might benefit from special services, and evaluating students' progress against state standards of performance and proficiency. Teachers also use assessment activities and results to inform students about teacher expectations. In other words, the assessment process can provide students with information about the kind of performance that they need to be successful in a given classroom and grade. Tests become a critical link in teaching when teachers provide students with clear feedback about results. Assessments likewise provide teachers valuable feedback about how successful they have been in achieving their instructional objectives and thereby help them chart the sequence and pace of future instructional activities.

Students also are decision makers, and they use classroom assessment information to influence many of their decisions. For example, many students set personal academic expectations for themselves based on teachers' assessments of prior achievement. Feedback they receive from teachers about their performances on classroom and standardized tests can directly affect students' decisions about their strengths and weaknesses, interests, study activities, and possible career plans.

The assessment activities and decisions of teachers affect parents as well as students. For example, many parents communicate educational and behavioral expectations to their children. Some parents also plan educational resources and establish home study environments to assist their children. Feedback from teachers about daily achievement, classroom tests, annual standardized tests, and statewide assessments often significantly influence parents' perceptions of their child and his or her teachers. Testing results also provide parents and others in the community with information about the school's performance. That is, does the school prepare students for the basic skills of reading, writing, and calculating? In summary, results from assessments of children's learning can significantly influence parents' attitudes about their children and schooling.

Clearly, the enterprise of assessing students often is very important in the lives of teachers, students, and many parents. Recognizing this, a joint committee on testing practices from major educational and psychological organizations developed a *Code of Fair Testing Practices in Education* (American Educational Research Association, 1988). This code contains standards for educational test developers and users in four areas: developing/selecting tests, interpreting scores, striving for fairness, and informing test takers. The code is meant for use by the general public and is included in its entirety as Appendix F in this book. With its focus on fairness and appropriate interpretation of test scores, this code serves as an appropriate conclusion to this book on educational assessment and the inclusion of all students in assessment programs.

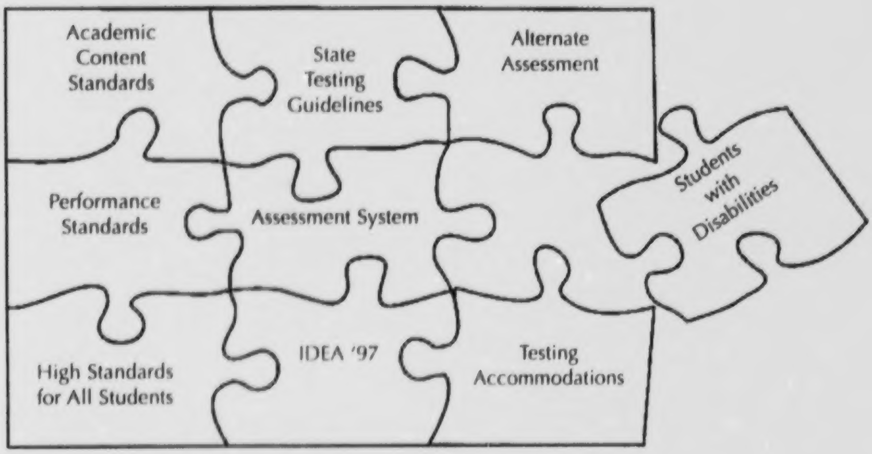


FIGURE 5.1
The Completed Educational Accountability Puzzle

Note. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 85. Copyright February 2000 Wisconsin Department of Public Instruction.

Completing the Cases and the Educational Accountability Puzzle

As stressed throughout this book, assessing all students is an important and, at times, challenging undertaking that requires knowledge of testing practices, test content, legal guidelines, and technical aspects of tests, as well as a clear understanding of students' learning objectives and instructional programs. We have highlighted these challenges throughout this book via the cases of Patrick, Tia, and Chris, so as we close this chapter we must also close the cases of these three students (see Case Study Conclusions). We hope it is clear that all of these students count in the big picture of their schools' accountability and all of them can meaningfully participate in assessments that communicate to their teachers, parents, and others about their many abilities.

If educators across the United States are going to actualize the requirements of IDEA '97 and the potential of standards-based education for *all* students, then *all* educators will need to have a strong understanding of their state's standards, the content of the tests covered in their state assessment, their state's testing guidelines, the valid use of testing accommodations, the valid use of alternate assessments, and how to communicate these assessment results to students and their families. As indicated early in this book, there are at least nine pieces to the educational accountability puzzle in most states (see Figure 5.1). As a result of reading this book and talking with colleagues about assessment activities like those required in large-scale assessments, we hope you are now prepared to facilitate the meaningful participation of all students in statewide and districtwide assessments. If so, you understand how the pieces to the accountability puzzle fit together!

Case Study Conclusions

Patrick

■ Patrick is a 9-year-old 4th-grader who has difficulty reading. Patrick lives with his parents in Florida and will soon be taking the state-mandated test for all 4th-graders. His parents were very concerned that Patrick would struggle on the test because of his reading problems. Although his reading has been somewhat delayed and is causing him some frustrations with learning, he does not qualify as a student with a disability. Consequently, in Florida he is ineligible for any testing accommodations and because he is functioning in the regular curriculum, an alternate assessment is not a reasonable tactic either.

Patrick's teacher has had several years of experience giving the state assessment and has a good understanding of the state's academic standards. In preparation for the statewide assessment, she has already planned some test preparation activities for her entire class each week for the month preceding the test. Patrick will benefit from these activities and will be better prepared to take the test, although his results in mathematics, social studies, and science may not be highly valid because of his reading difficulties.

Tia

■ Tia is an 8th-grader who is classified as learning disabled primarily in the area of reading. She receives all of her instruction in regular classes with some support from a consulting special education teacher. Her instructional reading level is approximately 5th grade, but she has good listening and memory skills and is a highly motivated student.

Her IEP team has communicated with her and her parents about the forthcoming test given by the Wisconsin Department of Public Instruction as part of its accountability system. The IEP team members have noted that Tia is eligible for testing accommodations and have developed a detailed accommodation plan based on their knowledge of her instructional accommodations and the procedure and responses requirements of TerraNova.

Tia participated in the test and received accommodations for her reading disability on all sections of the test except the reading subsection. For that portion, she received reading assistance only with a few words in the test directions. Her participation on the test was uneventful, and the resulting scores were deemed to be valid indicators of her abilities.

Chris

■ Chris is an 11th-grader diagnosed with Down syndrome He loves school and receives all of his instruction in a special education class with several other students and two teachers. His IEP focuses on functional living skills and emphasizes reading survival words and math skills associated with telling time and making change. He also has IEP objectives concerning social behavior related to job skills. Given that his curriculum is very different from the mainstream curriculum and he requires extensive instructional support, Chris qualified for an alternate assessment in Idaho.

The IAA utilizes a behavior rating scale approach that focuses on the collection of evidence for IEP objectives that are judged to be aligned with state academic standards in the content areas of reading, writing, and mathematics. Chris's teacher and teacher aide collected work samples from the classroom that were recent and representative of his typical effort. They then used a scoring guide and independently evaluated the evidence. They concluded, for state accountability purposes, that Chris was functioning at the Prerequisite Skills level in each of the content areas. For Chris's parents, they provided a detailed description of his strengths and weaknesses and provided a progress report for each of his IEP objectives. The results of the assessment led to a discussion of ways to revise Chris's IEP for the next year.

APPENDIX A

Standards for Teacher Competence in Educational Assessment of Students

Developed by the American Federation of Teachers
National Council on Measurement in Education
National Education Association

The professional education associations began working in 1987 to develop standards for teacher competence in student assessment out of concern that the potential educational benefits of student assessments be fully realized. The Committee[1] appointed to this project completed its work in 1990 following reviews of earlier drafts by members of the measurement, teaching, and teacher preparation and certification communities. Parallel committees of affected associations are encouraged to develop similar statements of qualifications for school administrators, counselors, testing directors, supervisors, and other educators in the near future. These statements are intended to guide the preservice and inservice preparation of educators, the accreditation of preparation programs, and the future certification of all educators.

A standard is defined here as a principle generally accepted by the professional associations responsible for the document. Assessment is defined as the process of obtaining information that is used to make educational decisions about students, to give feedback to the student about his or her progress, strengths, and weaknesses, to judge instructional effectiveness and curricular adequacy, and to inform policy. The various assessment techniques include, but are not limited to, formal and informal observation, qualitative analysis of pupil performance and products, paper-and-pencil tests, oral questioning, and analysis of student records. The assessment competencies included here are the knowledge and skills critical to a teacher's role as educator. It is understood that there are many competencies beyond assessment competencies which teachers must possess.

---

[1] The Committee that developed this statement was appointed by the collaborating professional associations: James R. Sanders (Western Michigan University) chaired the Committee and represented NCME along with John R. Hills (Florida State University) and Anthony J. Nitko (University of Pittsburgh). Jack C. Merwin (University of Minnesota) represented the American Association of Colleges for Teacher Education, Carolyn Trice represented the American Federation of Teachers, and Marcella Dianda and Jeffrey Schneider represented the National Education Association.

By establishing standards for teacher competence in student assessment, the associations subscribe to the view that student assessment is an essential part of teaching and that good teaching cannot exist without good student assessment. Training to develop the competencies covered in the standards should be an integral part of preservice preparation. Further, such assessment training should be widely available to practicing teachers through staff development programs at the district and building levels.

The standards are intended for use as:

- a guide for teacher educators as they design and approve programs for teacher preparation
- a self-assessment guide for teachers in identifying their needs for professional development in student assessment
- a guide for workshop instructors as they design professional development experiences for in-service teachers
- an impetus for educational measurement specialists and teacher trainers to conceptualize student assessment and teacher training in student assessment more broadly than has been the case in the past.

The standards should be incorporated into future teacher training and certification programs. Teachers who have not had the preparation these standards imply should have the opportunity and support to develop these competencies before the standards enter into the evaluation of these teachers.

The Approach Used to Develop the Standards

The members of the associations that supported this work are professional educators involved in teaching, teacher education, and student assessment. Members of these associations are concerned about the inadequacy with which teachers are prepared for assessing the educational progress of their students, and thus sought to address this concern effectively. A committee named by the associations first met in September 1987 and affirmed its commitment to defining standards for teacher preparation in student assessment. The committee then undertook a review of the research literature to identify needs in student assessment, current levels of teacher training in student assessment, areas of teacher activities requiring competence in using assessments, and current levels of teacher competence in student assessment.

The members of the committee used their collective experience and expertise to formulate and then revise statements of important assessment competencies. Drafts of these competencies went through several revisions by the Committee before the standards were released for public review.

Comments by reviewers from each of the associations were then used to prepare a final statement.

The Scope of a Teacher's Professional Role and Responsibilities for Student Assessment

There are seven standards in this document. In recognizing the need to revitalize classroom assessment, some standards focus on classroom-based competencies. Because of teachers' growing roles in education and policy decisions beyond the classroom, other standards address assessment competencies underlying teacher participation in decisions related to assessment at the school, district, state, and national levels.

The scope of a teacher's professional role and responsibilities for student assessment may be described in terms of the following activities. These activities imply that teachers need competence in student assessment and sufficient time and resources to complete them in a professional manner.

Activities Occurring Prior to Instruction

(a) Understanding students' cultural backgrounds, interests, skills, and abilities as they apply across a range of learning domains and/or subject areas; (b) understanding students' motivations and their interests in specific class content; (c) clarifying and articulating the performance outcomes expected of pupils; and (d) planning instruction for individuals or groups of students.

Activities Occurring During Instruction

(a) Monitoring pupil progress toward instructional goals; (b) identifying gains and difficulties pupils are experiencing in learning and performing; (c) adjusting instruction; (d) giving contingent, specific, and credible praise and feedback; (e) motivating students to learn; and (f) judging the extent of pupil attainment of instructional outcomes.

Activities Occurring After the Appropriate Instructional Segment (e.g., lesson, class, semester, grade)

(a) Describing the extent to which each pupil has attained both short- and long-term instructional goals; (b) communicating strengths and weaknesses based on assessment results to students, and parents or guardians; (c) recording and reporting assessment results for school-level analysis, evaluation, and decision-making; (d) analyzing assessment information gathered before and during instruction to understand each students' progress to date and to inform future instruction planning; (e) evaluating the effectiveness of instruction; and (f) evaluating the effectiveness of the curriculum and materials in use.

Activities Associated With a Teacher's Involvement in School Building and School District Decision-Making

(a) Serving on a school or district committee examining the school's and district's strengths and weaknesses in the development of its students; (b) working on the development or selection of assessment methods for school building or school district use; (c) evaluating school district curriculum; and (d) other related activities.

Activities Associated With a Teacher's Involvement in a Wider Community of Education

(a) Serving on a state committee asked to develop learning goals and associated assessment methods; (b) participating in reviews of the appropriateness of district, state, or national student goals and associated assessment methods and (c) interpreting the results of state and national student assessment programs.

Each standard that follows is an expectation for assessment knowledge or skill that a teacher should possess in order to perform well in the five areas just described. As a set, the standards call on teachers to demonstrate skill at selecting, developing, applying, using, communicating, and evaluating student assessment information and student assessment practices. A brief rationale and illustrative behaviors follow each standard.

The standards represent a conceptual framework or scaffolding from which specific skills can be derived. Work to make these standards operational will be needed even after they have been published. It is also expected that experience in the application of these standards should lead to their improvement and further development.

Standards for Teacher Competence in Educational Assessment of Students

1. **Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.**

   Skills in choosing appropriate, useful, administratively convenient, technically adequate, and fair assessment methods are prerequisite to good use of information to support instructional decisions. Teachers need to be well-acquainted with the kinds of information provided by a broad range of assessment alternatives and their strengths and weaknesses. In particular, they should be familiar with criteria for evaluating and selecting assessment methods in light of instructional plans.

   Teachers who meet this standard will have the conceptual and application skills that follow. They will be able to use the concepts of assessment error and validity when developing or selecting their approaches to classroom assessment of students. They will understand how valid assessment data can support instructional activities such as providing appropriate feedback to students, diagnosing group and individual learning needs, planning for individualized educational programs, motivating students, and evaluating instructional procedures. They will understand how invalid information can affect instructional decisions about students. They will be able to use and evaluate assessment options available to them, considering among other things, the cultural, social, economic, and language backgrounds of students. They will be aware that different assessment approaches can be incompatible with certain instructional goals and may impact quite differently on their teaching.

Teachers will know, for each assessment approach they use, its appropriateness for making decisions about their pupils. Moreover, teachers will know of where to find information about and/or reviews of various assessment methods. Assessment options are diverse and include text- and curriculum-embedded questions and test, standardized criterion-referenced and norm-referenced tests, oral questioning, spontaneous and structured performance assessments, portfolios, exhibitions, demonstrations, rating scales, writing samples, paper-and pencil tests, seatwork and homework, peer- and self-assessments, student records, observations, questionnaires, interviews, projects, products, and others' opinions.

2. **Teachers should be skilled in developing assessment methods appropriate for instructional decisions.**

While teachers often use published or other external assessment tools, the bulk of the assessment information they use for decision-making comes from approaches they create and implement. Indeed, the assessment demands of the classroom go well beyond readily available instruments.

Teachers who meet this standard will have the conceptual and application skills that follow. Teachers will be skilled in planning the collection of information that facilitates the decisions they will make. They will know and follow appropriate principles for developing and using assessment methods in their teaching, avoiding common pitfalls in student assessment. Such techniques may include several of the options listed at the end of the first standard. The teacher will select the techniques which are appropriate to the intent of the teacher's instruction.

Teachers meeting this standard will also be skilled in using student data to analyze the quality of each assessment technique they use. Since most teachers do not have access to assessment specialists, they must be prepared to do these analyses themselves.

3. **The teacher should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.**

It is not enough that teachers are able to select and develop good assessment methods; they must also be able to apply them properly. Teachers should be skilled in administering, scoring, and interpreting results form diverse assessment methods.

Teachers who meet this standard will have the conceptual and application skills that follow. They will be skilled in interpreting informal and formal teacher-produced assessment results, including pupils' performances in class and on homework assignments. Teachers will be able to use guides for scoring essay questions and projects, stencils for scoring response-choice questions, and scales for rating performance assessments. They will be able to use these in ways that produce consistent results.

Teachers will be able to administer standardized achievement tests and be able to interpret the commonly reported scores: percentile ranks, percentile band scores, standard scores, and grade equivalents. They will have a conceptual understanding of the summary indexes commonly reported with assessment results: measures of central tendency, dispersion, relationships, reliability, and errors of measurement.

Teachers will be able to apply these concepts of score and summary indices in ways that enhance their use of the assessments that they develop. They will be able to analyze assessment results to identify pupils' strengths and errors. If they get inconsistent results, they will seek other explanations for the discrepancy or other data to attempt to resolve the uncertainty before arriving at a decision. They will be able to use assessment methods in ways that encourage students' educational development and that do not inappropriately increase students' anxiety levels.

4. **Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.**

Assessment results are used to make educational decisions at several levels: in the classroom about students, in the community about a school and a school district, and in society, generally, about the purposes and outcomes of the educational enterprise. Teachers play a vital role when participating in decision-making at each of these levels and must be able to use assessment results effectively.

Teachers who meet this standard will have the conceptual and application skills that follow. They will be able to use accumulated assessment information to organize a sound instructional plan for facilitating students' educational development. When using assessment results to plan and/or evaluate instruction and curriculum, teachers will interpret the results correctly and avoid common misinterpretations, such as basing decisions on scores that lack curriculum validity. They will be informed about the results of local, regional, state, and national assessments and about their appropriate use for pupil, classroom, school, district, state, and national educational improvement.

5. **Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.**

Grading students is an important part of professional practice for teachers. Grading is defined as indicating both a student's level of performance and a teacher's valuing of that performance. The principles for using assessments to obtain valid grades are known and teachers should employ them.

Teachers who meet this standard will have the conceptual and application skills that follow. They will be able to devise, implement, and explain a procedure for developing grades composed of marks from various assignments, projects, in-class activities, quizzes, tests, and/or other assessments that they may use. Teachers will understand and be able to articulate why the grades they assign are rational, justified, and fair, acknowledging that such grades reflect their preferences and judgments. Teachers will be able to recognize and to avoid faulty grading procedures such as using grades as punishment. They will be able to evaluate and to modify their grading procedures in order to improve the validity of the interpretations made from them about students' attainments.

6. **Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.**

Teachers must routinely report assessment results to students and to parents or guardians. In addition, they are frequently asked to report or to discuss assessment results with other educators and with diverse lay audiences. If the results are not communicated effectively, they may be misused or not used. To communicate effectively with others on matters of student assessment, teachers must be able to use assessment terminology appropriately and must be able to articulate the meaning, limitations, and implications of assessment procedures and their interpretations of them. At other times, teachers may need to help the public to interpret assessment results appropriately.

Teachers who meet this standard will have the conceptual and application skills that follow. Teachers will understand and be able to give appropriate explanations of how the interpretation of student assessments must be moderated by the student's socio-economic, cultural, language, and other background factors. Teachers will be able to explain that assessment results do not imply that such background factors limit a student's ultimate educational development. They will be able to communicate to students and to their parents or guardians how they may assess the student's educational progress. Teachers will understand and be able to explain the importance of taking measurement errors into account when using assessments to make decisions about individual students. Teachers will be able to explain the limitations of different informal and formal assessment methods. They will be able to explain printed reports of the results of pupil assessments at the classroom, school district, state, and national levels.

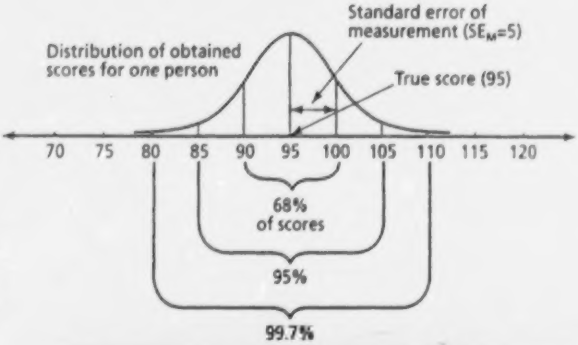
7. **Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.**

Fairness, the rights of all concerned, and professional ethical behavior must undergird all student assessment activities, from the initial planning for and gathering of information to the interpretation, use, and communication of the results. Teachers must be well-versed in their own ethical and legal responsibilities in assessment. In addition, they should also attempt to have the inappropriate assessment practices of others discontinued whenever they are encountered. Teachers should also participate with the wider educational community in defining the limits of appropriate professional behavior in assessment.

Teachers who meet this standard will have the conceptual and application skills that follow. They will know those laws and case decisions which affect their classroom, school district, and state assessment practices. Teachers will be aware that various assessment procedures can be misused or overused resulting in harmful consequences such as embarrassing student, violating a student's right to confidentiality, and inappropriately using students' standardized achievement test scores to measure teaching effectiveness.

Calculating the Standard Error of Measurement

Hypothetical Distribution Illustrating the Standard Error of Measurement


Standard error of measurement (SE$_M$=5)
Distribution of obtained scores for one person
True score (95)
70
75
80
85
90
95
100
105
110
115
120
68% of scores
95%
99.7%

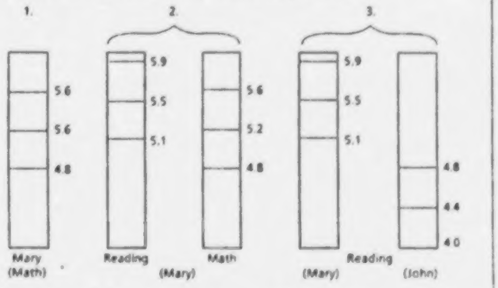
Theoretical Explanation of the Standard Error of Measurement

1. **It is assumed that each person has a *true score* on a particular test, a hypothetical value representing a score free of error (true score + 95 on the diagram).**
2. **If a person could be tested repeatedly (without memory, practice effects, or other changes), the average of the obtained scores would *be approximately normally distributed* around the true score (see diagram).**
3. **From what is known about the normal distribution curve, approximately 68 percent of the obtained scores would fall within one standard error of measurement of the person's true score; approximately 95 percent of the scores would fall within two standard errors; and approximately 99.7 percent of the scores would fall within three standard errors.**
4. **Although the true score can never be known, the standard error of measurement can be applied to a person's obtained score to set "reasonable limits" for locating the true score (e.g., an obtained score of 97 ± 5 = 92 to 102).**
5. **These "reasonable limits" provide *confidence bands* for interpreting an obtained score. When the standard error of measurement is small, the confidence band is narrow (indicating high reliability), and thus we have greater confidence that the obtained score is near the true score.**

GUIDELINES

Practical Applications of the Standard Error of Measurement in Test Interpretation

A confidence band one standard error above and below the obtained score is commonly used in test profiles to aid in interpreting individual scores and in judging whether differences between scores are likely to be "real differences" or differences caused by chance.



1. *Interpreting an individual score.* The confidence band indicates "reasonable limits" within which to locate the true score (Mary's math score probably falls somewhere between 4.8 and 5.6).
2. *Interpreting the difference between two scores from a test battery.* When the ends of the bands overlap, there is no "real difference" between scores (Mary's scores in reading and math show no meaningful difference).
3. *Interpreting the difference between the scores of two individuals on the same test.* When ends of bands do not overlap, there is a "real difference" between scores (Mary's reading score is higher than John's).

Note. From Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability for all students: Facilitating the meaningful participation of students with disabilities in district and statewide assessment programs.* Madison: Wisconsin Department of Public Instruction, p. 85. Copyright February 2000 Wisconsin Department of Public Instruction.

APPENDIX C

Iowa Tests of Basic Skills

Overview
Scores
Test Forms Available
Score & Test Characteristics
Unique Features & Publisher Information
Individual Performance Profile
Student Criterion-Referenced Skills Analysis
Mathematics Sample Items
Social Studies Sample Items

Iowa Tests of Basic Skills
Overview

Norm Sample	1992 Forms K & L: 136,934 Students 1995 Form M: 126,468 Students • Separate norms available for large city, Catholic, private, high socioeconomic status, low socioeconomic status, and international schools. • Generally representative of U.S. Census Data
Content Areas	• Core Battery: Listening (Levels 5–8 only), Word Analysis (Levels 5–8 only), Vocabulary, Reading, Language, and Mathematics. • Complete Battery (adds beginning at level 7) Social Studies, Science, and Sources of Information. • Survey Battery: Reading, Language, and Mathematics (uses subsets of items from the Core Battery). • Writing Assessment and a Listening Assessment available (Levels 9 through 14). • Constructed response supplements (form 1) available for levels 9–14.
Population	• Grades K–12

Stated Purpose	To assess the "basic skills" or "the entire range of skills a student needs to progress satisfactorily through school" Including: higher-order thinking skills, interpretation, classification, comparison, analysis, and inference.
Scores	Vocabulary, Listening, Language, Language Total, Mathematics, Core Total, Word Analysis (optional), Mathematics Advanced Skills, Mathematics Total, Reading Advanced Skills, Reading Total, Reading, Listening Language, Mathematics Concepts, Mathematics Problems, Mathematics Computation (optional), Social Studies, Science, Sources of Information, Composite, Language Advanced Skills, Mathematics Advanced Skills, Survey Battery Total, Reading Comprehension, Spelling, Capitalization, Punctuation, Usage and Expression, Mathematics Concepts and Estimation, Mathematics Problem Solving and Data Interpretation, Mathematics Total, Maps and Diagrams, Reference Materials, Sources of Information Total, Composite
Support Materials	Practice tests for each level, teacher manuals for test preparation and directions for administration and scoring, interpretive guides for administrators, teachers and counselors, content outline; rater training material and anchor papers for written tests.

Iowa Tests of Basic Skills
Scores

Criterion-Referenced Scores	Range	Description
Student Criterion-Referenced Skill Analysis	0–100	• Report for each skill area on the number and percentage of items the student answered correctly along with the average percent-correct score for class and nation • Based on the number right scores provided for the subskills measured within each test
Norm Referenced Scores	**Range**	**Description**
Developmental Scale Scores	0–999	• Describes performance across grades on the same scale • Generally comparable across test/grade level • Provided only for composite scores (not available for subdomain scores) • Good for longitudinal evaluation studies and evaluating subject area programs over time
Percentiles	1–99	• Proportion of students in the same grade whose score is less than or equal to the score • Individual and School Percentiles Available
Normal Curve Equivalent (NCE)	1–99	• Explains the status of score relative to a normal distribution in equal interval units
Grade Equivalent	K–12	• Typical score for a student in the identified grade and month of school
Stanine	1–9	• Same as Normal Curve Equivalent, but uses larger, less precise intervals

Iowa Tests of Basic Skills
Test Forms Available

Form	Target Grade Level	Core Battery	Complete Battery	Survey Battery	Writing Assessment	Listening Assessment	Constructed Response Supplement
Level 5	K.1–1.5	✓	✓*	✓	☒	☒	☒
Level 6	K.8–1.9	✓	✓*	✓	☒	☒	☒
Level 7	1.7–2.6	✓	✓	✓	☒	☒	☒
Level 8	2.5–3.5	✓	✓	✓	☒	☒	☒
Level 9	3	✓	✓	✓	✓	✓	✓
Level 10	4	✓	✓	✓	✓	✓	✓
Level 11	5	✓	✓	✓	✓	✓	✓
Level 12	6	✓	✓	✓	✓	✓	✓
Level 13	7	✓	✓	✓	✓	✓	✓
Level 14	8	✓	✓	✓	✓	✓	✓

- Indicates no difference from Core Battery in terms of subdomains tested
- ✓ Available
- ☒ Not Available

Notes	• Part of Riverside's Integrated Assessment System; • Co-normed with Iowa Tests of Educational Development, the Tests of Achievement, Proficiency and the Cognitive Abilities test, and the Performance Assessments for ITBS.

Iowa Tests of Basic Skills
Score & Test Characteristics

	Number of Items Per Test	Test Time (Hrs: Min.)	Reliability	Constructed Response	Selected Response
Complete Battery	146–515	2:10–3:31	+ .90	0	All Administered
Survey Battery	114–142	2:10–5:26	+.90	0	All Administered
Listening	29–31	25–30	.67–.79	0	All Administered
Word Analysis	30–38	15–20	.80–.90	0	All Administered
Vocabulary	20–42	15–20	.80–.90	0	All Administered
Reading	34–62	43–55	.80–.90	0	All Administered
Language	29–153	25–1:06	.80–.90	0	All Administered
Mathematics	29–162	25–1:10	.80–.90	0	All Administered
Social Studies	31–43	25–30	.80–.90	0	All Administered

Science	31–43	25–30	.59 and above	0	All Administered
Sources of Information	22–69	30–55	.80–.90	0	All Administered
Writing Assessment **(Narrative, Explanation, Description, Informative Report, and persuasive)**	Varies	Varies	Reader reliability (holistic & analytic scoring) .80 and below Student .50	All Administered	0

Note: Although most subtests of Forms K, L, and M are in the .80s and .90s, Levels 5–8 have lower reliabilities (~ .80s). Higher levels have higher reliabilities (~.90). Thus, the above reliabilities represent the *range* of reliability scores reported and may not apply to a particular form or level. Times and numbers of items reflect entire domain. Times derived for Levels 5–8 reflect approximations—they are not timed.

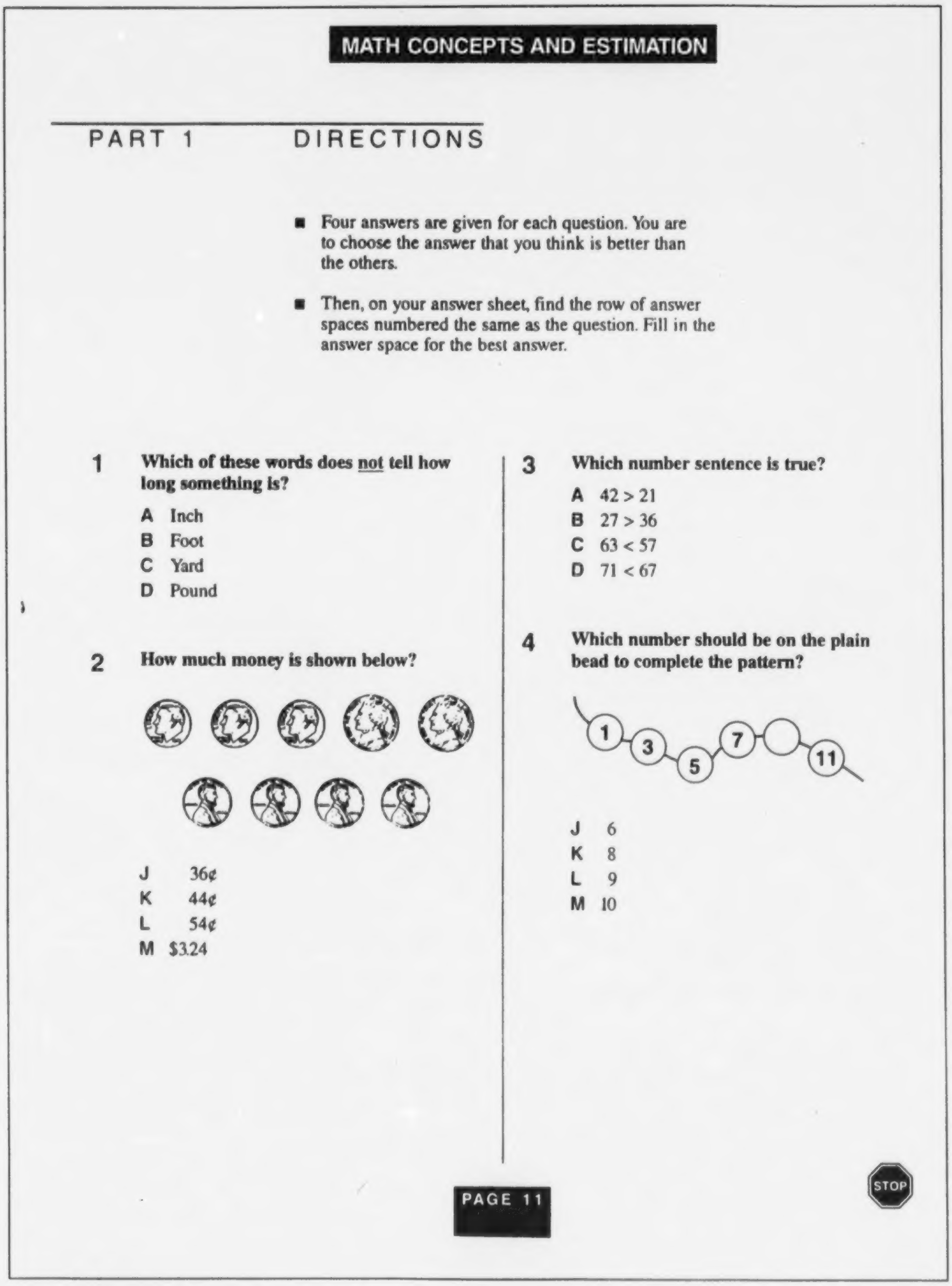
Constructed Response Items Available as a supplement

Standards Alignment	National Council of Teachers of Mathematics National Council of Teachers of English Curriculum and Evaluation Standards for School Mathematics

Iowa Tests of Basic Skills
Unique Features

Directions for accommodations	• Yes • Extensive research based directions for accommodating test takers available from publisher
Inclusion of Students with Disabilities in Normative Sample	• Yes
Inclusion of Accommodations in the Standardization Procedures	• Yes
Clear Specification of Access Skills for each Test	• No
Availability of large print and Braille forms	• Yes • Available from publisher
Availability of locator test for out of level testing	• Yes • Available from publisher
Availability of test practice materials	• Yes • Available from publisher

Availability of Multiple Forms	• Yes • Available from publisher
Publisher Information	• The Iowa Tests of Basic Skills is published by Riverside Publishing For More Information Contact: Riverside Publishing 425 Spring Lake Drive Itasca IL 60143-9921 Tel: 1-800-323-9540 Fax: 630-467-7192 For Specific Information Relevant to your State or find the regional office nearest you, visit the Riverside website at: http://www.riverpub.com
** Note: As is true of most comprehensive achievement tests, the Iowa Tests of Basic Skills can be purchased in many versions. Also, some states will contract with the publisher to customize the test for a particular state. Consequently, the information contained in this appendix cannot be applied to other forms of the Iowa Tests of Basic Skills. Although this information should be generally accurate for most forms of the Complete Battery of the ITBS, this information may not apply to *specific* versions. Consult your school district or state assessment coordinator for specific details regarding the version your school/district uses.	



MATH CONCEPTS AND ESTIMATION

PART 1 DIRECTIONS

- Four answers are given for each question. You are to choose the answer that you think is better than the others.
- Then, on your answer sheet, find the row of answer spaces numbered the same as the question. Fill in the answer space for the best answer.

1 **Which of these words does not tell how long something is?**

A Inch
B Foot
C Yard
D Pound

2 **How much money is shown below?**

J 36¢
K 44¢
L 54¢
M $3.24

3 **Which number sentence is true?**

A $42 > 21$
B $27 > 36$
C $63 < 57$
D $71 < 67$

4 **Which number should be on the plain bead to complete the pattern?**

J 6
K 8
L 9
M 10

Mathematics Sample Items

Note. From Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., Dunbar, S. B., Oberley, K. R., Bray, G. B., Lewis, J. C., & Qualls, A. L. (1996). ITBS Form M Practice Tests. Chicago: Riverside, p. 11. Copyright © 1996 by The University of Iowa. Reprinted by permission.

SOCIAL STUDIES

Directions: Questions 4 through 6 are based on the information below.

At the time the French Revolution began, capitalism was rising and France was going through an era of rapid economic growth. The glitter and luxury of the court life of the French king were the envy of the German and Italian princes. The reputations of French writers and philosophers had strengthened the country's position as a world leader in the Century of Enlightenment. France had a prosperous agriculture and a large merchant marine. In addition, the Industrial Revolution was taking root on her soil.

In the midst of this good fortune, however, there were serious problems that demanded solutions. The crucial problem was that rapid economic and intellectual development were not matched by necessary social and political change.

The population was divided into three groups, or Estates. The First Estate consisted of the clergy and officials of the church. Although it included only one-half of 1 percent of a French population of 25 million, this group controlled 20 percent of the land and had an income one-half that of the king.

The nobles, about 110,000 in number, made up the Second Estate. They were not subject to most of the heavy taxes. The best positions in the government were open only to them, and their large holdings of farmland provided them a handsome income.

The rest of the population made up the Third Estate. Most of its members were peasants and farm laborers who were still living close to the poverty level. The remainder of the Third Estate was the middle class, or bourgeoisie, which was the main force behind the commercial growth of the country. This new class of merchants and professionals was angry because it did not have political and social influence equal to its economic position. When the bourgeoisie would no longer accept this imbalance between its economic strength and its political weakness, the Revolution began.

4 **The pre-Revolutionary tax system in France can be described as**

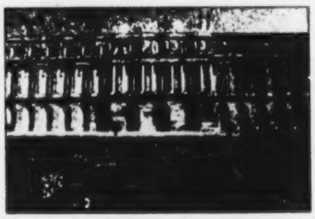
- **J** discriminatory.
- **K** democratic.
- **L** progressive.
- **M** illegal.

5 **The author traces the origin of the French Revolution to causes which were essentially**

- **A** international in nature.
- **B** demographic in nature.
- **C** religious in nature.
- **D** political in nature.

6 **Which of these statements is best supported by the passage?**

- **J** The middle class was opposed to the Industrial Revolution.
- **K** Changes in a society's economic structure often result in political changes.
- **L** Democracy is always necessary for economic growth.
- **M** All wars have economic causes.



STOP

PAGE 15

Social Studies Sample Items

Note. From Feldt, L. S., Forsyth, R. A., Ansley, T. N., & Alnot, S. D. (1996). ITED® Form M Practice Tests. Chicago: Riverside, p. 15. Copyright © 1996 by The University of Iowa. All rights reserved. Reprinted by permission.

BEST COPY AVAILABLE

Stanford Achievement Test (9th Edition)

Overview
Scores
Test Forms Available
Score & Test Characteristics
Unique Features & Publisher Information
Student Report for Multiple-Choice Battery
Group Report for Multiple-Choice Battery
Reading Sample Items
Mathematics Sample Items

Stanford Achievement Test 9th Edition
Overview

Norm Sample	Fall 1995: 200,000 K–12 Spring 1995: 250,000 K–12 • Generally representative of U.S. Census data • Different norms available for Catholic, Private, High SES, and Urban Schools
Content Areas	Reading, Language, Spelling, Study Skills, Listening, Mathematics, Science, Social Science, Writing
Population	K–12
Scores	Reading, Language, Spelling, Study Skills, Listening, Mathematics, Science, Social Science, Writing
Stated Purpose	To "measure student achievement in reading, language, spelling, study skills, listening, mathematics, science, and social science."
Support Materials	Teachers' guide to administration, Student Answer Sheets (primary 3 only) Practice tests for each level, teacher manuals for test preparation, and directions for administration and scoring.

Stanford Achievement Test 9th Edition Scores

Criterion-Referenced Scores	**Range**	**Description**
Objective Referenced Information	1–3	• Sub-domain scores can be reported as either content or process clusters • Available as raw scores, number of items attempted, and number of items possible
Performance Standards	1–4	• Proficiency scores relative to curricular expectations
	Performance Level Categories * available as either 3 or 4 levels of achievement	
	Score	**Description**
	1	• *Below Basic* • Less than partial mastery
	2	• *Basic* • Partial Mastery
	3	• *Proficient* • Solid academic performance
	4	• *Advanced* • Superior performance

Norm Referenced Scores	Range	Description
Scale Scores		• Equivalent across forms and levels of the same subtest and domain total
Individual and Group Percentiles	1–99	• Proportion of students in the same grade whose score is less than or equal to the score
Individual Normal Curve Equivalent (NCE)	1–99	• Explains the status of score relative to a normal distribution in equal interval units
Individual and Group Stanine	1–9	• Same as Normal Curve Equivalent, but uses larger intervals
Individual Grade Equivalents	1–12	• Grade level for whom the score was 'average' or 'typical.' (e.g., a grade equivalent of 4.5 is the average score among children in the 6th month of fourth grade). • Grade equivalents cannot be compared across subjects (i.e. the jump from 3rd to 5th grade level work in Language may not reflect the same amount of learning as the jump from 3rd to 5th grade in Mathematics).
Individual Achievement/Ability Comparisons	High Middle Low	• Discrepancies between performance and ability as compared to other students of same ability level • Addresses "Whether or not child is learning to the best of his/her ability?" • Must also administer Otis - Lennon School Ability Test 7th edition to obtain scores

	Score	**Description**
	High	• Score among highest of scores made by students of same measured ability
	Middle	• Score was in the middle of scores made by students of same measured ability
	Low	• Score was among lowest of scores made by students of same measured ability
Objective-referenced Information		• Available as content clusters for each subskill within a subtest or domain • Available as process clusters representing thinking and reasoning in each content area
Content Clusters		• Looks at specific cognitive skills within a general subject area and identifies specific strengths and/or areas of difficulty
	Score	**Description**
	Above Average	• Score was in the top 23%
	Average	• Score was in the middle 54%
	Low	• Score was in the bottom 23%
Open-Ended Assessment Performance Indicators	0–3	• Measure performance on Open-Ended Assessment Supplement of Test Only • Available as item, cluster, and total scores

Stanford Achievement Test 9th Edition
Test Forms Available

Form	Target Grade Level	Basic Battery	Complete Battery	Abbreviated Battery	Open Ended Battery
Stanford Early Achievement Test 1 (SESAT 1)	K.0 – K.5	✓	✓	☒	☒
SESAT 2	K.5 – 1.5	✓	✓	☒	☒
Primary I	1.5 – 2.5	✓	✓	✓	✓
Primary 2	2.5 – 3.5	✓	✓	✓	✓
Primary 3	3.5 – 4.5	✓	✓	✓	✓
Intermediate 1	4.5 – 5.5	✓	✓	✓	✓
Intermediate 2	5.5 – 6.5	✓	✓	✓	✓
Intermediate 3	6.5 – 7.5	✓	✓	✓	✓
Advanced 1	7.5 – 8.5	✓	✓	✓	✓

Advanced 2	8.5 – 9.9	✓	✓	✓	✓
Stanford Test of Academic Skills (Task 1)	9.0 – 9.9	✓	✓	✓	✓
Task 2	10.0 – 10.9	✓	✓	✓	✓
Task 3	11.0 – 13.0	✓	✓	✓	✓

- ✓ Available
- ☒ Not Available

Stanford Achievement Test 9th Edition
Score & Test Characteristics

	Number of Items	Test Time (Hrs: Min.)	Reliability	Constructed Response	Selected Response
Complete Battery	198–392	2:15–3:45	+.90	0	All Administered
Basic Battery	158–312	1:45–4:35	+.90	0	All Administered
Abbreviated Battery	110–150	1:35–2:20	+.90	0	All Administered
Language (form SA)	46–54	40 min–45	Composite	All Administered	0
Reading	78–110	45–1:25	+.90	0	All Administered
Mathematics	40–82	30–1:20	+.90	0	All Administered
Language (form S)	44–48	40–45	+.90	0	All Administered
Spelling	30	20–25	.70 and below	0	All Administered
Study Skills	30	20–25	.70 and below	0	All Administered
Listening	40	30	.70 and below	0	All Administered

Social Science	40	20–25	.70 and below	0	All Administered
Science	40	20–25	.70 and below	0	All Administered
Notes	• Various methods of estimating reliability will produce different scores. The above estimates were made using an alternate forms method. • IF combined with open ended section reliability rates raise to .80–.90 • Inter-rater coefficients for open ended writing assessments ranged from .50–mid .80 (Pearson correlation) to .70–.90 (Spearman brown) • Reliability of performance scales: .61–.88 • Spelling and Study skills not included as separate subtest when alternate language (form SA) subtest is used				
Standards Alignment	National Assessment of Educational Progress, Curriculum and Evaluation Standards for School Mathematics, Writing Process Model, Science for All Americans, National Science Education Standards, Scope, Sequence and Coordination of Secondary Science Project, Benchmark for Science Literacy, Bradley Commission on History in the Schools, CIVITAS, National Standards for Civics and Government, GENIP Standards, Geography Education Standards Project, National Council for the Social Studies Curriculum Standards, Joint Council on Economic Education.				

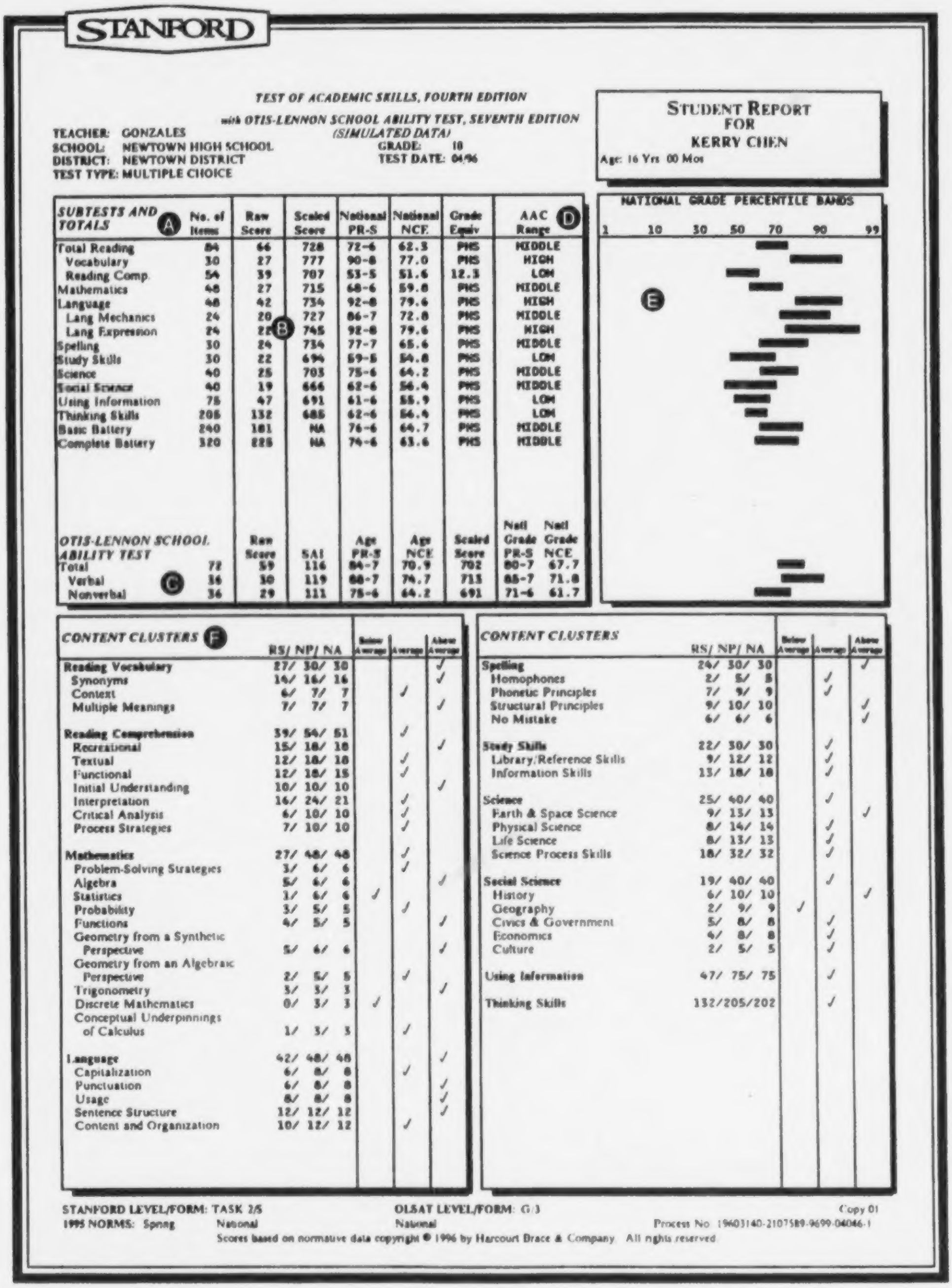
Stanford Achievement Test 9th Edition
Unique Features

Directions for accommodations	• Yes • Available from publisher
Inclusion of Students with Disabilities in Normative Sample	• Yes
Inclusion of Accommodations in the Standardization Procedures	• Yes
Clear Specification of Access Skills for each Test	• Yes • Available in pamphlet form from publisher
Availability of large print and Braille forms	• Yes • Available from publisher
Availability of locator test for out of level testing	• Yes • Available from publisher
Availability of test practice materials	• Yes • Available from publisher

Availability of Multiple forms	• Yes • Available from publisher
Notes	• Can be integrated with Otis-Lennon School Ability Test For Achievement/Ability Comparisons • Variable scoring or partial credit options available on open-ended tests
Publisher Information	• The Stanford Achievement Test 9th Edition is published by Harcourt Brace Educational Measurement. For More Information Contact: Harcourt Brace Educational Measurement 555 Academic Court San Antonio TX 78204-2498 Tel: 1-800-211-8378 Fax: 1-800-232-1223 E-Mail: Customer_Service@harcourt.com For Specific Information Relevant to your State or to find the regional office nearest you, visit the Harcourt website at: http://www.hemweb.com/index.htm

** Note: As is true of most comprehensive achievement tests, the Stanford Achievement Test 9th Edition can be purchased in many versions. Also, some states will contract with the publisher to customize the test for a particular state. Consequently, the information contained in this appendix applied to the Multiple and Abbreviated Multiple form of the Stanford 9. Although this information should be generally accurate for most forms of the Stanford 9, this information may not apply to *specific* versions. Consult your school district or state assessment coordinator for specific details regarding the version your school/district uses.

Exhibit A page 1 of 5


NATIONAL GRADE PERCENTILE BANDS
1
10
30
50
70
90
99
Ⓔ

STANFORD

TEST OF ACADEMIC SKILLS, FOURTH EDITION
with OTIS-LENNON SCHOOL ABILITY TEST, SEVENTH EDITION
(SIMULATED DATA)

TEACHER: GONZALES
SCHOOL: NEWTOWN HIGH SCHOOL
DISTRICT: NEWTOWN DISTRICT
TEST TYPE: MULTIPLE CHOICE

GRADE: 10
TEST DATE: 04/96

STUDENT REPORT
FOR
KERRY CHEN
Age: 16 Yrs 00 Mos

SUBTESTS AND TOTALS Ⓐ	No. of Items	Raw Score	Scaled Score	National PR-S	National NCE	Grade Equiv	AAC Range Ⓓ
Total Reading	84	66	728	72-6	62.3	PHS	MIDDLE
Vocabulary	30	27	777	90-8	77.0	PHS	HIGH
Reading Comp.	54	39	707	53-5	51.6	12.3	LOW
Mathematics	48	27	715	68-6	59.8	PHS	MIDDLE
Language	48	42	734	92-8	79.6	PHS	HIGH
Lang Mechanics	24	20	727	86-7	72.8	PHS	MIDDLE
Lang Expression	24	22 Ⓑ	745	92-8	79.6	PHS	HIGH
Spelling	30	24	734	77-7	65.6	PHS	MIDDLE
Study Skills	30	22	694	59-5	54.8	PHS	LOW
Science	40	25	703	75-6	64.2	PHS	MIDDLE
Social Science	40	19	666	62-6	56.4	PHS	MIDDLE
Using Information	75	47	691	61-6	55.9	PHS	LOW
Thinking Skills	205	132	686	62-6	56.4	PHS	LOW
Basic Battery	240	181	NA	76-6	64.7	PHS	MIDDLE
Complete Battery	320	225	NA	74-6	63.6	PHS	MIDDLE

OTIS-LENNON SCHOOL ABILITY TEST		Raw Score	SAI	Age PR-S	Age NCE	Scaled Score	Natl Grade PR-S	Natl Grade NCE
Total	72	59	116	84-7	70.9	702	80-7	67.7
Verbal Ⓒ	36	30	119	88-7	74.7	713	85-7	71.8
Nonverbal	36	29	111	75-6	64.2	691	71-6	61.7

CONTENT CLUSTERS Ⓕ	RS/ NP/ NA	Below Average	Average	Above Average
Reading Vocabulary	27/ 30/ 30			✓
Synonyms	14/ 16/ 16			✓
Context	6/ 7/ 7		✓	
Multiple Meanings	7/ 7/ 7			✓
Reading Comprehension	39/ 54/ 51		✓	
Recreational	15/ 18/ 18			✓
Textual	12/ 18/ 18		✓	
Functional	12/ 18/ 15		✓	
Initial Understanding	10/ 10/ 10			✓
Interpretation	16/ 24/ 21		✓	
Critical Analysis	6/ 10/ 10		✓	
Process Strategies	7/ 10/ 10		✓	
Mathematics	27/ 48/ 48		✓	
Problem-Solving Strategies	3/ 6/ 6		✓	
Algebra	5/ 6/ 6			✓
Statistics	1/ 6/ 6	✓		
Probability	3/ 5/ 5		✓	
Functions	4/ 5/ 5			✓
Geometry from a Synthetic Perspective	5/ 6/ 6			✓
Geometry from an Algebraic Perspective	2/ 5/ 5		✓	
Trigonometry	3/ 3/ 3			✓
Discrete Mathematics	0/ 3/ 3	✓		
Conceptual Underpinnings of Calculus	1/ 3/ 3		✓	
Language	42/ 48/ 48			✓
Capitalization	6/ 8/ 8		✓	
Punctuation	6/ 8/ 8			✓
Usage	8/ 8/ 8			✓
Sentence Structure	12/ 12/ 12			✓
Content and Organization	10/ 12/ 12		✓	

CONTENT CLUSTERS	RS/ NP/ NA	Below Average	Average	Above Average
Spelling	24/ 30/ 30			✓
Homophones	2/ 5/ 5		✓	
Phonetic Principles	7/ 9/ 9		✓	
Structural Principles	9/ 10/ 10			✓
No Mistake	6/ 6/ 6			✓
Study Skills	22/ 30/ 30		✓	
Library/Reference Skills	9/ 12/ 12		✓	
Information Skills	13/ 18/ 18		✓	
Science	25/ 40/ 40		✓	
Earth & Space Science	9/ 13/ 13			✓
Physical Science	8/ 14/ 14		✓	
Life Science	8/ 13/ 13		✓	
Science Process Skills	18/ 32/ 32		✓	
Social Science	19/ 40/ 40		✓	
History	6/ 10/ 10			✓
Geography	2/ 9/ 9	✓		
Civics & Government	5/ 8/ 8		✓	
Economics	4/ 8/ 8		✓	
Culture	2/ 5/ 5		✓	
Using Information	47/ 75/ 75		✓	
Thinking Skills	132/205/202		✓	

STANFORD LEVEL/FORM: TASK 2/5
1995 NORMS: Spring National

OLSAT LEVEL/FORM: G/3
National

Copy 01
Process No. 19603140-2107589-9699-04046-1

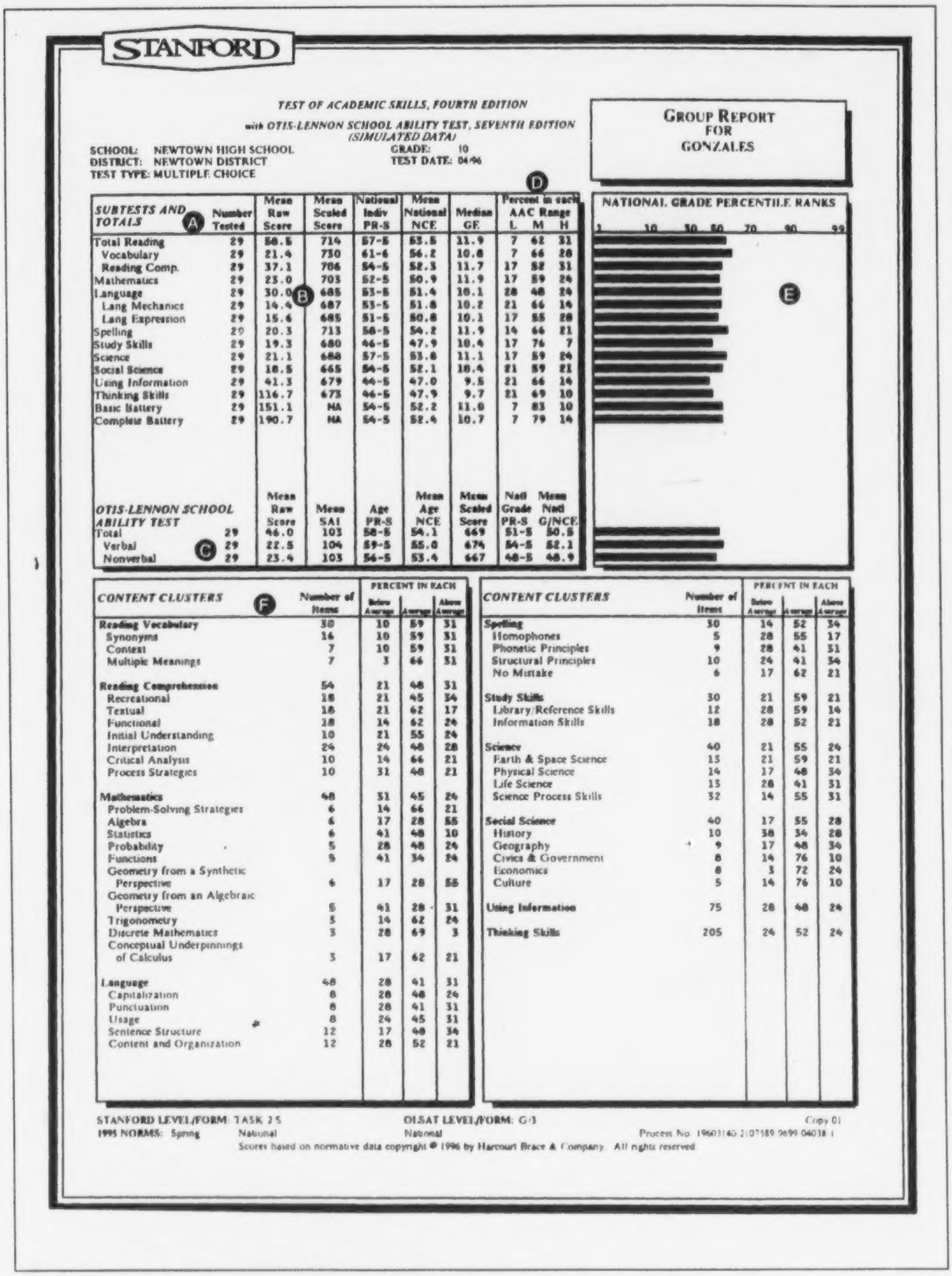
Scores based on normative data copyright © 1996 by Harcourt Brace & Company. All rights reserved.

Student Report for Multiple Choice Battery

Note. From the Guide for Classroom Planning: Task 1/2/3 of the *Stanford Achievement Test: Ninth Edition*. Copyright © 1996 by Harcourt, Inc. Reproduced by permission. All rights reserved.

BEST COPY AVAILABLE

Exhibit A page 2 of 5



STANFORD

TEST OF ACADEMIC SKILLS, FOURTH EDITION
with OTIS-LENNON SCHOOL ABILITY TEST, SEVENTH EDITION
(SIMULATED DATA)

GROUP REPORT
FOR
GONZALES

SCHOOL: NEWTOWN HIGH SCHOOL
DISTRICT: NEWTOWN DISTRICT
TEST TYPE: MULTIPLE CHOICE
GRADE: 10
TEST DATE: 04/96

SUBTESTS AND TOTALS Ⓐ	Number Tested	Mean Raw Score	Mean Scaled Score	National Indiv PR-S	Mean National NCE	Median GE	Percent in each AAC Range Ⓓ L	M	H
Total Reading	29	58.5	714	57-5	53.5	11.9	7	62	31
Vocabulary	29	21.4	730	61-6	56.2	10.8	7	66	28
Reading Comp.	29	37.1	706	54-5	52.3	11.7	17	52	31
Mathematics	29	23.0	703	52-5	50.9	11.9	17	59	24
Language	29	30.0 Ⓑ	685	53-5	51.4	10.1	28	48	24
Lang Mechanics	29	14.4	687	53-5	51.8	10.2	21	66	14
Lang Expression	29	15.6	685	51-5	50.8	10.1	17	55	28
Spelling	29	20.3	713	58-5	54.2	11.9	14	66	21
Study Skills	29	19.3	680	46-5	47.9	10.4	17	76	7
Science	29	21.1	688	57-5	53.8	11.1	17	59	24
Social Science	29	18.5	665	54-5	52.1	10.4	21	59	21
Using Information	29	41.3	679	44-5	47.0	9.5	21	66	14
Thinking Skills	29	116.7	673	46-5	47.9	9.7	21	69	10
Basic Battery	29	151.1	NA	54-5	52.2	11.0	7	83	10
Complete Battery	29	190.7	NA	54-5	52.4	10.7	7	79	14

OTIS-LENNON SCHOOL ABILITY TEST		Mean Raw Score	Mean SAI	Age PR-S	Mean Age NCE	Mean Scaled Score	Nat'l Grade PR-S	Mean Nat'l G/NCE
Total	29	46.0	103	58-5	54.1	669	51-5	50.5
Verbal Ⓒ	29	22.5	104	59-5	55.0	674	54-5	52.1
Nonverbal	29	23.4	103	56-5	53.4	667	48-5	48.9

NATIONAL GRADE PERCENTILE RANKS

1 10 30 50 70 90 99

Ⓔ

CONTENT CLUSTERS Ⓕ	Number of Items	PERCENT IN EACH Below Average	Average	Above Average
Reading Vocabulary	30	10	59	31
Synonyms	16	10	59	31
Context	7	10	59	31
Multiple Meanings	7	3	66	31
Reading Comprehension	54	21	48	31
Recreational	18	21	45	34
Textual	18	21	62	17
Functional	18	14	62	24
Initial Understanding	10	21	55	24
Interpretation	24	24	48	28
Critical Analysis	10	14	66	21
Process Strategies	10	31	48	21
Mathematics	48	31	45	24
Problem-Solving Strategies	6	14	66	21
Algebra	6	17	28	55
Statistics	6	41	48	10
Probability	5	28	48	24
Functions	5	41	34	24
Geometry from a Synthetic Perspective	6	17	28	55
Geometry from an Algebraic Perspective	5	41	28	31
Trigonometry	5	14	62	24
Discrete Mathematics	3	28	69	3
Conceptual Underpinnings of Calculus	3	17	62	21
Language	48	28	41	31
Capitalization	8	28	48	24
Punctuation	8	28	41	31
Usage	8	24	45	31
Sentence Structure	12	17	48	34
Content and Organization	12	28	52	21

CONTENT CLUSTERS	Number of Items	PERCENT IN EACH Below Average	Average	Above Average
Spelling	30	14	52	34
Homophones	5	28	55	17
Phonetic Principles	9	28	41	31
Structural Principles	10	24	41	34
No Mistake	6	17	62	21
Study Skills	30	21	59	21
Library/Reference Skills	12	28	59	14
Information Skills	18	28	52	21
Science	40	21	55	24
Earth & Space Science	13	21	59	21
Physical Science	14	17	48	34
Life Science	13	28	41	31
Science Process Skills	32	14	55	31
Social Science	40	17	55	28
History	10	38	34	28
Geography	9	17	48	34
Civics & Government	8	14	76	10
Economics	8	3	72	24
Culture	5	14	76	10
Using Information	75	28	48	24
Thinking Skills	205	24	52	24

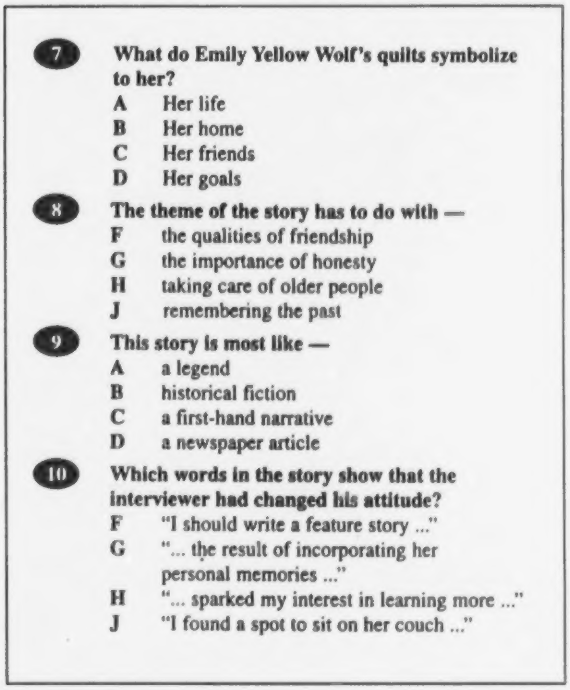
STANFORD LEVEL/FORM: TASK 2/S　OLSAT LEVEL/FORM: G/3　Copy 01
1995 NORMS: Spring　National　National　Process No. 19603140-2107589-9699-04038-1
Scores based on normative data copyright © 1996 by Harcourt Brace & Company. All rights reserved.

Group Report for Multiple-Choice Battery

Note. From the Guide for Classroom Planning: Task 1/2/3 of the *Stanford Achievement Test: Ninth Edition.* Copyright © 1996 by Harcourt, Inc. Reproduced by permission. All rights reserved.

BEST COPY AVAILABLE

Exhibit A page 3 of 5



7 **What do Emily Yellow Wolf's quilts symbolize to her?**
- A Her life
- B Her home
- C Her friends
- D Her goals

8 **The theme of the story has to do with —**
- F the qualities of friendship
- G the importance of honesty
- H taking care of older people
- J remembering the past

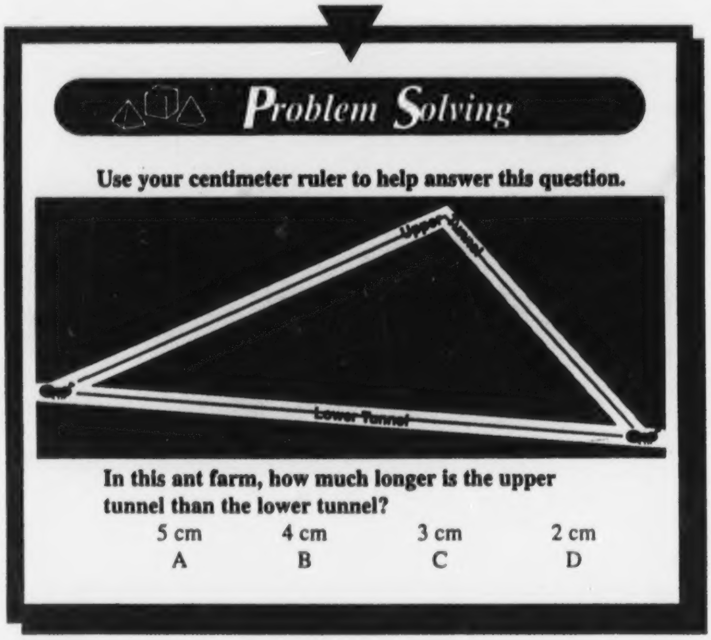
9 **This story is most like —**
- A a legend
- B historical fiction
- C a first-hand narrative
- D a newspaper article

10 **Which words in the story show that the interviewer had changed his attitude?**
- F "I should write a feature story ..."
- G "... the result of incorporating her personal memories ..."
- H "... sparked my interest in learning more ..."
- J "I found a spot to sit on her couch ..."

Reading Sample Items

Note. From the Stanford 9 Special Report Reading of the *Stanford Achievement Test: Ninth Edition.* Copyright © 2000 by Harcourt, Inc. Reproduced by permission. All rights reserved.

Exhibit A page 4 of 5



Upper Tunnel
Lower Tunnel

Problem Solving

Use your centimeter ruler to help answer this question.

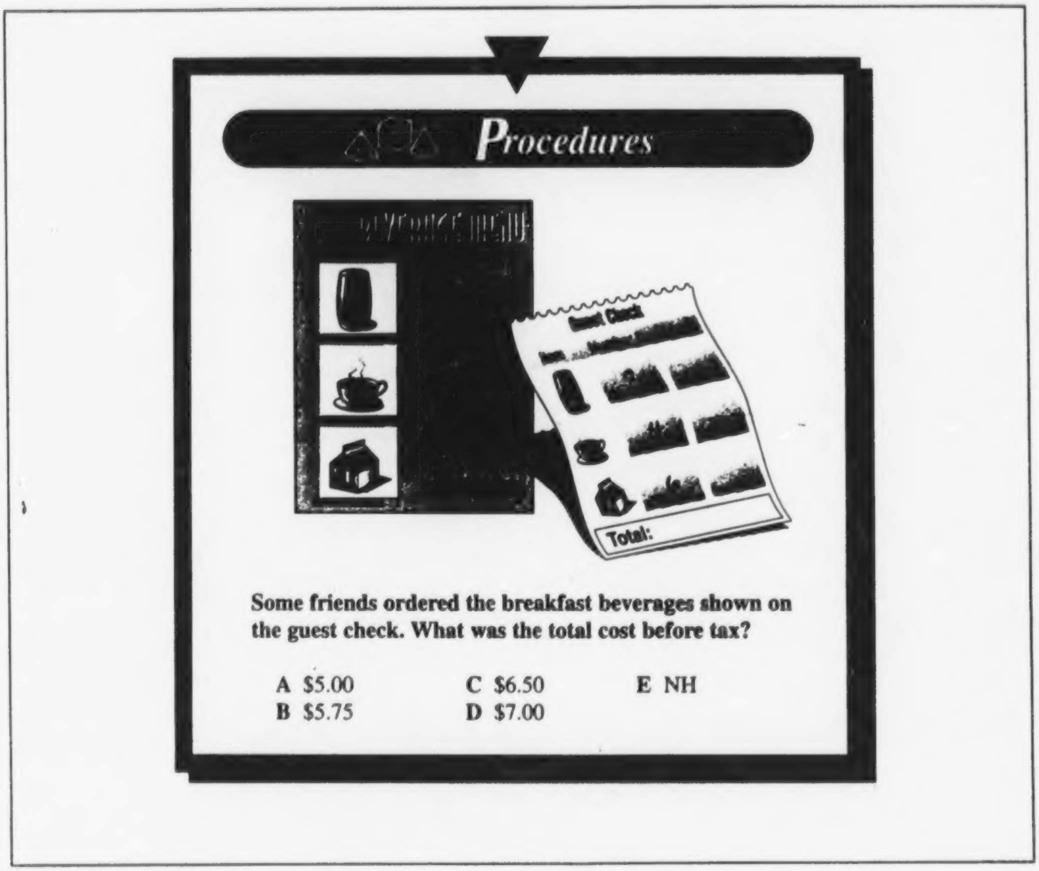
In this ant farm, how much longer is the upper tunnel than the lower tunnel?

5 cm	4 cm	3 cm	2 cm
A	B	C	D

Mathematics Sample Items

Note. From the Stanford 9 Special Report Mathematics of the *Stanford Achievement Test: Ninth Edition*. Copyright © 2000 by Harcourt, Inc.. Reproduced by permission. All rights reserved.

Exhibit A page 5 of 5


Procedures
Total:

Some friends ordered the breakfast beverages shown on the guest check. What was the total cost before tax?

A $5.00
B $5.75
C $6.50
D $7.00
E NH

Mathematics Sample Items

Note. From the Stanford 9 Special Report Mathematics of the *Stanford Achievement Test: Ninth Edition.* Copyright © 2000 by Harcourt, Inc.. Reproduced by permission. All rights reserved.

BEST COPY AVAILABLE

APPENDIX E

TerraNova

Overview
Scores
Test Forms Available
Score & Test Characteristics
Unique Features & Publisher Information
Individual Profile Report
Reading Sample Items
Mathematics Sample Items

TerraNova
Overview

Norm Sample	Fall 1996: 71,366 Students Grades 1–12 Spring 1997: 100,000 Students Grades: K–12 * Generally representative of U.S. Census data
Content Areas	Reading, Language Arts, Mathematics, Social Studies, Science
Population	Grades K–12 LE–10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21–22.
Scores	Reading, Language Arts, Mathematics, Science, Social Studies, Word Analysis, Vocabulary, Language Mechanics, Spelling, Mathematics Computation, Communication Arts.
Stated Purpose	To provide a "comprehensive, modular assessment series of student achievement"
Support Materials	Teachers' guide to administration, interpretation, and using test results. Practice tests for each level, teacher manuals for test preparation, and directions for administration and scoring.

TerraNova Scores

Criterion-Referenced Scores	Range	Description
Developmental Scale Scores	0–999	IRT score with Equal Intervals. • Generally comparable across test/grade level. • Provided only for composite scores (not available for subdomain scores) • Good for longitudinal evaluation studies and evaluating subject area programs over time
Objective Performance Index	0–100	Estimated number of items that the student would get correct if there had been 100 items measuring that objective
	Objective Performance Mastery Categories	
	Score	**Description**
	0–49	*Non-Mastery.* The student has not mastered the objective
	50–74	*Partial Mastery* The student lacks complete mastery or is inconsistent in performing the objective
	75–100	*Mastery* The student has mastered the objective and performs it consistently

Performance Levels	1–5	• Proficiency scores relative to curricular expectations Grades 1–2 (Primary) Grades 3–5 (Elementary) Grades 6–8 (Middle School) Grades 9–12 (High School) • All levels are interpreted relative to curricular expectations most appropriate for the student's current grade placement. • Similar to state and NAEP proficiency levels; but not linked to specific state academic standards or state-specific curricular expectations.
	Performance Level Categories	
	Score	**Description**
	1	• Starting out. • Very Limited Proficiency; Little command of domain
	2	• Progressing • Beginning or early mastery of domain.
	3	• Nearing Proficiency • Somewhat below expected Mastery
	4	• Proficient • Meeting Curricular expectations
	5	• Advanced • Exceeding curricular expectations

Norm Referenced Scores	Range	Description
Percentiles	1–99	Proportion of students in the same grade whose score is less than or equal to the score
Normal Curve Equivalent (NCE)	1–99	Explains the status of score relative to a normal distribution in equal interval units
Stanine	1–9	Same as Normal Curve Equivalent, but uses larger intervals

TerraNova
Test Forms Available

Form	Target Grade Level	Survey & Survey Plus	Complete Battery & Complete Battery Plus	Basic & Basic Plus	Multiple Assessments
Form 10	K.6–1.6	☒	✓ *	☒	☒
Form 11	1.6–2.6	☒	✓ **	✓ **	✓
Form 12	2.0–3.2	✓	✓	✓	✓
Form 13	2.6–4.2	✓	✓	✓	✓
Form 14	3.6–5.2	✓	✓	✓	✓
Form 15	4.6–6.2	✓	✓	✓	✓
Form 16	5.6–7.2	✓	✓	✓	✓
Form 17	6.6–8.2	✓	✓	✓	✓
Form 18	7.6–9.2	✓	✓	✓	✓
Form 19	8.6–10.2	✓	✓	✓	✓
Form 20	9.6–11.2	✓	✓	✓	✓
Form 21/22	10.6–12.9	✓	✓	✓	✓

* Science and Social Studies not available ** No language or spelling section in Plus Supplement Word Analysis subsections not available for forms 15–21/22	
Notes	• Part of Riverside's Integrated Assessment System • Instruments may be administered alone or in any combination • Available in English and in Spanish editions

TerraNova
Score & Test Characteristics

	Reading	Language Arts	Mathematics	Science	Social Studies
Number of Items Per Domain	43–48	23–26	35–43	35	34
Number of Items Per Test	62–64		35–43	35	34
Constructed Response	7–10		10–11	10	9
Selected Response	50–55		25–32	25	25
Test Time (Hrs: Min.)	1: 45		1:30	1:05	1:05
Standards Alignment	International Reading Association National Council of Teachers of English		National Council of Teachers of Mathematics	National Science Education Standards	** Multiple Sources

Objectives Tested	Basic Understanding Analyze Text Evaluate and Extend Meaning Identify Reading Strategies Intro to Print Sentence Structure Writing Strategies Editing Strategies		Number and Number Relations Computation and Estimation Operations Communication Concepts Problem Solving Measurement Geometry and Spatial Sense Data Analysis Statistics and Probability Patterns Functions	Science Inquiry Physical Science Life Science Earth & Space Science Science & Technology Personal and Social Perspectives in Science History and Nature of Science	Geographic Perspectives Historical and Cultural Perspectives Civics and Government Perspectives Economic Perspectives
Alpha	0.95	0.90	0.92	0.90	0.91

Sub-Domain Scores

Subdomain	**Reading**	**Vocabulary**	**Language**	**Mechanics**	**Mathematics**	**Computation**
Alpha	0.93	0.84	0.85	0.84	0.91	0.71

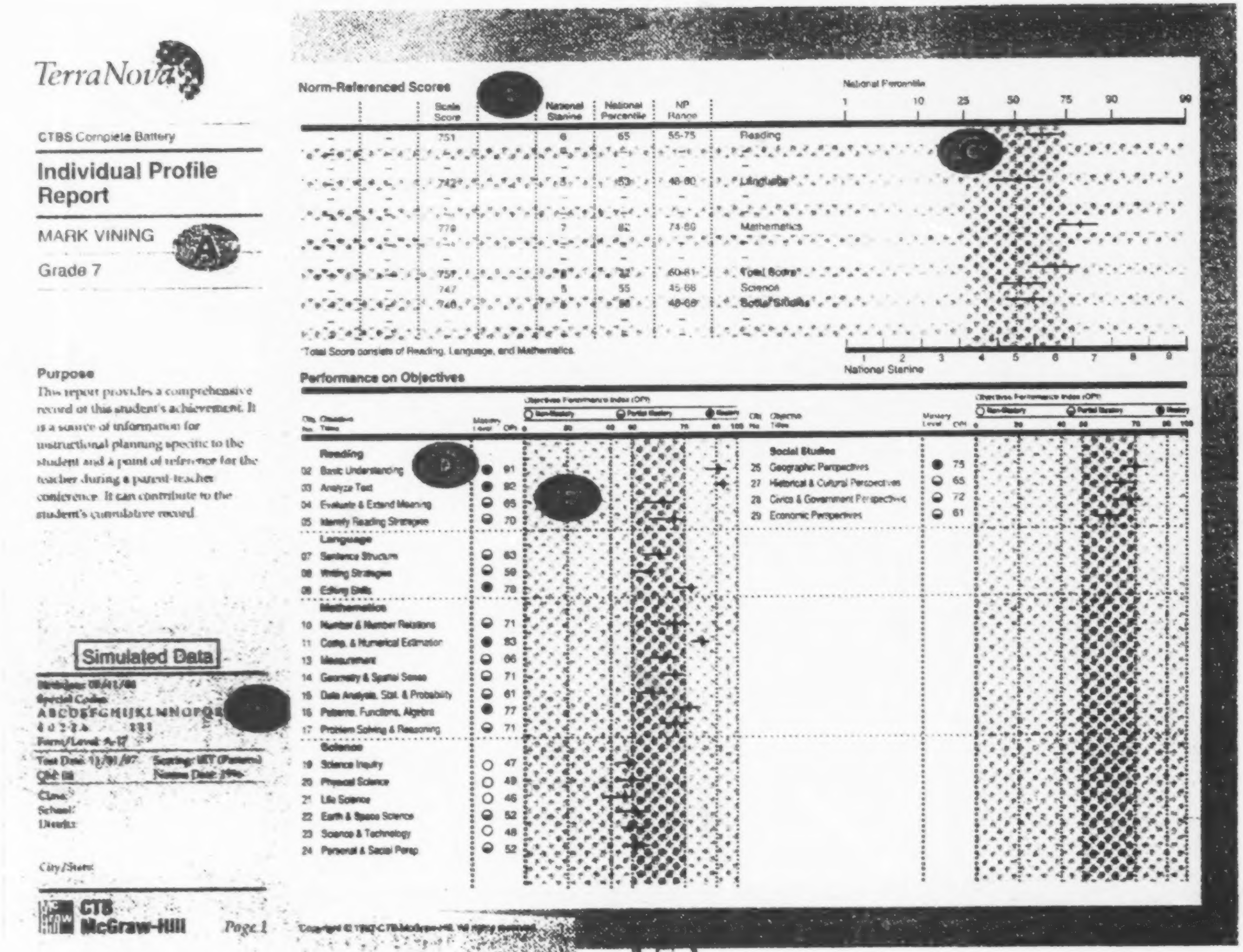
TerraNovã
Unique Features

Directions for accommodations	• Yes • For a general statement of accommodations see: http://www.ctb.com/ctb_features/10meeting.shtml • Specific directions for accommodations not available
Inclusion of Students with Disabilities in Normative Sample	• Yes
Inclusion of Accommodations in the Standardization Procedures	• Yes
Clear Specification of Access Skills for each Test	• Yes • Available from publisher
Availability of large print and Braille forms	• Yes • Available from publisher
Availability of locator test for out of level testing	• Yes • Available from publisher
Availability of test practice materials	• Yes • Recommended for one-time use only • May be purchased from publisher
Availability of Multiple forms	• Yes • May be purchased from publisher

<table>
<tr><td colspan="2">• Test items are grouped by theme and often follow realistic length of reading passages (e.g., a student reads a 400–1,200 word passage and then answers a series of questions about the passage)
• Item content is described with respect to curricular skills <i>and</i> cognitive skills
• Cognitive skill analysis based on 'Dimensions of Thinking' and classifies items by the types of cognitive skills demanded by the item.</td></tr>
<tr><td>Publisher Information</td><td>• The TerraNova is published by CTB-McGraw Hill and is based on the California Test of Basic Skills (CTBS).

<u>For More Information Contact:</u>
CTB–McGraw Hill
20 Ryan Ranch Road, Monterey, CA 93940
Tel: (800) 538-9547 Fax: (800) 282-0266

For specific information relevant to your State or to find the regional office nearest you, visit the CTB-McGraw Hill website at:
http://www.ctb.com</td></tr>
<tr><td colspan="2">** <u>Note</u>: As is true of most comprehensive achievement tests, the TerraNova can be purchased in many versions. Also, some states will contract with the publisher to customize the test for a particular state. Consequently, the information contained in this appendix applied to the Multiple Assessments (the most comprehensive) form of the Terra Nova. Although this information should be generally accurate for most forms of the TerraNova, this information may not apply to <i>specific</i> versions. Consult your school district or state assessment coordinator for specific details regarding the version your school/district uses.</td></tr>
</table>


National Percentile
1
10
25
50
75
90
99
National Stanine
1
2
3
4
5
6
7
8
9
Objectives Performance Index (OPI)
○ Non-Mastery
◒ Partial Mastery
● Mastery
0
25
50
75
100
A

TerraNova

CTBS Complete Battery

Individual Profile Report

MARK VINING

Grade 7

Purpose

This report provides a comprehensive record of this student's achievement. It is a source of information for instructional planning specific to the student and a point of reference for the teacher during a parent-teacher conference. It can contribute to the student's cumulative record.

Simulated Data

Birthdate: 08/41/86
Special Codes:
ABCDEFGHIJKLMNOPQR
4 0 2 2 A 331
Form/Level: A-17

Test Date: 11/01/97 Scoring: IRT (Pattern)
QM: 08 Norms Date: 1996

Class:
School:
District:

City/State:

CTB McGraw-Hill Page 1

Norm-Referenced Scores

		Scale Score		National Stanine	National Percentile	NP Range	
–	–	751		6	65	55-75	Reading
–	–	742		5	53	48-60	Language
–	–	779		7	82	74-89	Mathematics
–	–	757		6	72	60-81	Total Score*
–	–	747		5	55	45-66	Science
–	–	748		5	56	48-66	Social Studies

*Total Score consists of Reading, Language, and Mathematics.

Performance on Objectives

Obj. No.	Objective Titles	Mastery Level	OPI
	Reading		
02	Basic Understanding	●	91
03	Analyze Text	●	92
04	Evaluate & Extend Meaning	◒	65
05	Identify Reading Strategies	◒	70
	Language		
07	Sentence Structure	◒	63
08	Writing Strategies	◒	59
09	Editing Skills	●	78
	Mathematics		
10	Number & Number Relations	◒	71
11	Comp. & Numerical Estimation	●	83
13	Measurement	◒	66
14	Geometry & Spatial Sense	◒	71
15	Data Analysis, Stat. & Probability	◒	61
16	Patterns, Functions, Algebra	●	77
17	Problem Solving & Reasoning	◒	71
	Science		
19	Science Inquiry	○	47
20	Physical Science	○	49
21	Life Science	○	46
22	Earth & Space Science	◒	52
23	Science & Technology	○	48
24	Personal & Social Persp.	◒	52
	Social Studies		
26	Geographic Perspectives	●	75
27	Historical & Cultural Perspectives	◒	65
28	Civics & Government Perspective	◒	72
29	Economic Perspectives	◒	61

Copyright © 1997 CTB/McGraw-Hill. All rights reserved.

174

Individual Profile Report

174

BEST COPY AVAILABLE

Reading/Language Arts
Items Illustrating Content
Level 16

Read the passage. Then do Numbers 6 through 8.

Sojourner Truth

Chicago Historical Society, ICHI-22022

On a hot day in 1843, a thin, black woman wearing a gray dress, white turban, and sunbonnet left New York City. She left with a bag of clothes, twenty-five cents, and a new name. Born a slave named Isabella Baumfree, she had been freed in 1827 by New York State's Emancipation Act. As a slave, she had worked long, hard days in the fields. After being freed, she had been a house servant. She had helped slaves who escaped to the North find homes and jobs. Now she had changed her name to Sojourner Truth and was setting out to preach and sing about God, the evils of slavery, and the joy of being free.

Even though she could not read or write, Sojourner Truth was a powerful speaker. She was over six feet tall and had a booming voice and expressive, appealing eyes. Her simple words and songs attracted huge crowds. She influenced many people to join the fight against slavery.

When the Civil War began in 1861, Sojourner Truth raised money to buy gifts for Union soldiers by giving lectures and singing. She went into the camps and distributed the gifts herself. While traveling from one Union camp to another, she often gathered information about the Confederate troops, which she passed along to the next Union camp.

After the Civil War, Sojourner Truth continued her public speaking. Now, however, she spoke about women's rights, a cause she had worked for since she attended the first Women's Rights Convention in Worcester, Massachusetts, in 1850. She inspired women to work for the vote, equal pay, and equal rights under the law. At eighty years of age, ill health forced Sojourner Truth to give up her lecture tours. However, her message continued to inspire people everywhere.

Reading Sample Item

Reading/Language Arts
Items Illustrating Content
Level 16

6 **Which of these statements expresses an opinion?**

A Sojourner Truth was over six feet tall.
✓ **B** Sojourner Truth was a powerful speaker.
C Ill health forced Sojourner Truth to give up her lecture tours.
D Sojourner Truth left New York City with a bag of clothes, twenty-five cents, and a new name.

04 Evaluate and Extend Meaning

In this item, the student is asked to distinguish between fact and opinion. Other items in this objective focus on demonstrating an understanding of author's purpose, tone, bias, or point of view; extending and applying passage meaning to new situations, predicting future events or actions; and engaging in other types of critical assessment.

7 **In which section of a public library would you most likely find books to help you learn more about Sojourner Truth?**

A drama
B fiction
✓ **C** biography
D periodicals

05 Identify Reading Strategies

This item measures the student's ability to apply reading strategies by identifying genre criteria. Other items in this objective cover finding support for answers to linked items, formulating questions to explore deeper meaning, and summarizing text. Items in this objective can also ask the student to compare information across two passages or make connections between text and graphic representations of text information.

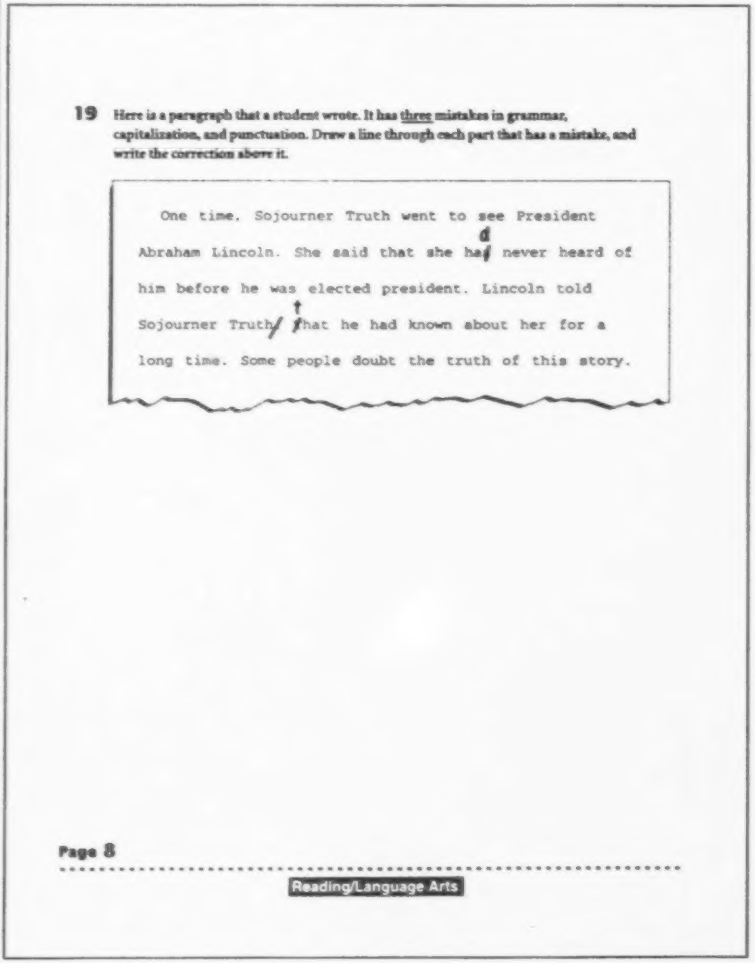
8 Choose the sentence that is complete and written correctly.

✓ **A** Born in New York in 1797, Isabella Baumfree grew up speaking only Dutch.
B The New York State Emancipation Act legally freed Isabella, she had run away six months earlier.
C Isabella having changed her name to Sojourner Truth in 1843.
D The story of Sojourner Truth's life, she dictated it to a friend, was published in 1850.

07 Sentence Structure

This item asks the student to identify correct sentence structure. Other items in this objective cover misplaced modifiers, sentence fragments, and sentence combining.

Reading Sample Item



19 Here is a paragraph that a student wrote. It has three mistakes in grammar, capitalization, and punctuation. Draw a line through each part that has a mistake, and write the correction above it.

One time. Sojourner Truth went to see President Abraham Lincoln. She said that she ha$\overset{d}{\cancel{s}}$ never heard of him before he was elected president. Lincoln told Sojourner Truth$\cancel{.}$ $\overset{t}{\cancel{T}}$hat he had known about her for a long time. Some people doubt the truth of this story.

Page 8

Reading/Language Arts

Reading/Language Arts Grade 6

09 Editing Skills

Item 19 **3 points**

- 1 point for changing *has* to *had*
- 1 point for eliminating the period after *Truth*
- 1 point for lowercasing the *T* in *That*

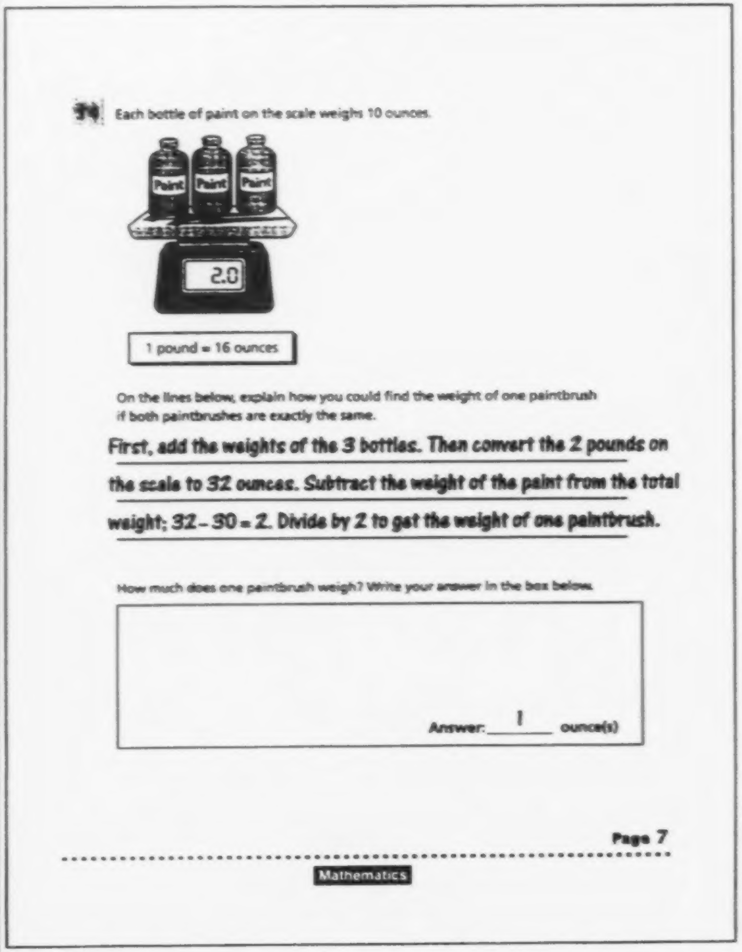
NOTE: If the student corrects mistakes in other acceptable ways, give full credit.

Mistakes need not be crossed out as long as they are corrected.

Ignore any changes made to parts of the sentences that do not have mistakes.

Misspellings of corrections are acceptable.

BEST COPY AVAILABLE



14 Each bottle of paint on the scale weighs 10 ounces.

Paint
Paint
Paint
2.0

1 pound = 16 ounces

On the lines below, explain how you could find the weight of one paintbrush if both paintbrushes are exactly the same.

First, add the weights of the 3 bottles. Then convert the 2 pounds on the scale to 32 ounces. Subtract the weight of the paint from the total weight; 32 - 30 = 2. Divide by 2 to get the weight of one paintbrush.

How much does one paintbrush weigh? Write your answer in the box below.

Answer: 1 ounce(s)

Page 7

Mathematics

Mathematics Grade 6

13 Measurement
17 Problem Solving and Reasoning

Item 14 **2 points**
- 1 point for correct explanation or mathematical process
- 1 point for answer of *1* (ounce)

Other acceptable response:

- 3 x 10 = 30

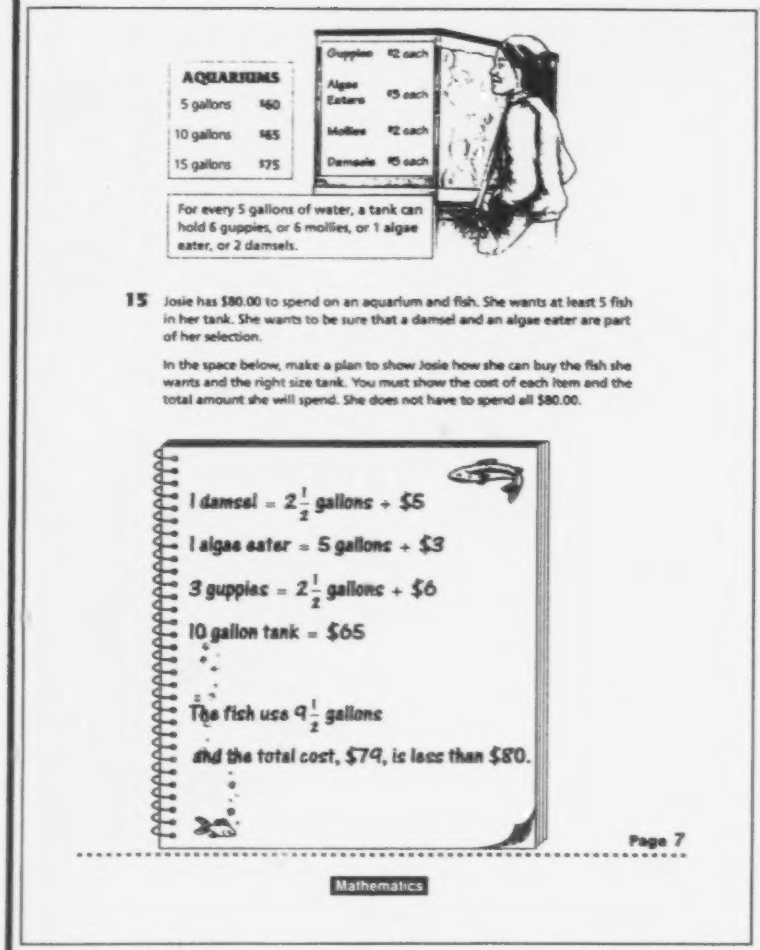
  16 x 2 = 32

  32 – 30 = 2

  2 ÷ 2 = 1

Math Sample Item

BEST COPY AVAILABLE



AQUARIUMS	
5 gallons	$60
10 gallons	$65
15 gallons	$75

Guppies	$2 each
Algae Eaters	$5 each
Mollies	$2 each
Damsels	$5 each

For every 5 gallons of water, a tank can hold 6 guppies, or 6 mollies, or 1 algae eater, or 2 damsels.

15 Josie has $80.00 to spend on an aquarium and fish. She wants at least 5 fish in her tank. She wants to be sure that a damsel and an algae eater are part of her selection.

In the space below, make a plan to show Josie how she can buy the fish she wants and the right size tank. You must show the cost of each item and the total amount she will spend. She does not have to spend all $80.00.

1 damsel = $2\frac{1}{2}$ gallons + $5

1 algae eater = 5 gallons + $3

3 guppies = $2\frac{1}{2}$ gallons + $6

10 gallon tank = $65

The fish use $9\frac{1}{2}$ gallons

and the total cost, $79, is less than $80.

Page 7

Mathematics

Mathematics Grade 7

16 Patterns, Functions, Algebra
17 Problem Solving and Reasoning

Item 15 **2 points**

- 1 point for a plan that has a total of $80.00 or less AND at least 5 fish
- 1 point for keeping within the gallon restraints for the fish chosen

BEST COPY AVAILABLE

Math Sample Item

APPENDIX F

Code of Fair Testing Practices in Education

Prepared by the Joint Committee on Testing Practices

The Code of Fair Testing Practices in Education states the major obligations to test takers of professionals who develop or use educational tests. The Code is meant to apply broadly to the use of tests in education (admissions, educational assessment, educational diagnosis, and student placement). The Code is not designed to cover employment testing, licensure or certification testing, or other types of educational tests, it is directed primarily at professionally developed tests such as those sold by commercial test publishers or used in formally administered programs. The Code is not intended to cover tests made by individual teachers for use in their own classrooms.

The Code addresses the roles of test developers and test users separately. Test users are people who select tests, commission test development services, or make decisions on the basis of test scores. Test developers are people who actually construct tests as well as those who set policies for particular testing programs. The roles may, of course, overlap as when a state education agency commissions test development process, and makes decisions on the basis of the test scores.

The Code presents standards for educational test developers and users in four areas:

A. Developing/Selecting Tests
B. Interpreting Scores
C. Striving for Fairness
D. Informing Test Takers

The Code has been developed by the Joint Committee on Testing Practices, a cooperative effort of several professional organizations, that has as its aim the advancement, in the public interest, of the quality of testing practices. The Joint Committee was initiated by the American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education. In addition to these three groups, the American Association for Counseling and Development/Association for Measurement and Evaluation in Counseling and Development, and the American Speech-Hearing Association are now also sponsors of the Joint Committee.

This is not copyrighted material. Reproduction and dissemination are encouraged. Please cite this document as follows:

Code of Fair Testing Practices in Education. (1988) Washington, D.C.: Joint Committee on Testing Practices. (Mailing Address: Joint Committee on Testing Practices, American Psychological Association, 750 First Avenue, NE, Washington, D.C., 20002-4242.

Organizations, institutions, and individual professionals who endorse the Code commit themselves to safeguarding the rights of test takers by following the principles listed. The Code is intended to be consistent with the relevant parts of the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985). However, the Code differs from the Standards in both audience and purpose. The Code is meant to be understood by the general public; it is limited to educational tests; and the primary focus is on those issues that affect the proper use of tests. The Code is not meant to add new principles over and above those in the Standards or to change the meaning of the Standards. The goal is rather to represent the spirit of a selected portion of the Standards in a way that is meaningful to test takers and/or their parents or guardians. It is the hope of the Joint Committee that the Code will also be judged to be consistent with existing codes of conduct and standards of other professional groups who use educational tests.

A. Developing/Selecting Appropriate Tests*

Test developers should provide the information that test users need to select appropriate tests.

Test users should select tests that meet the purpose for which they are to be used and that are appropriate for the intended test-taking populations.

Test Developers Should:	**Test Users Should:**
1. Define what each test measures and what the test should be used for. Describe the population(s) for which the test is appropriate.	1. First define the purpose for testing and the population to be tested. Then, select a test for that purpose and that population based on a thorough review of the available information.
2. Accurately represent the characteristics, usefulness, and limitations of tests for their intended purposes.	2. Investigate potentially useful sources of information, in addition to test scores, to corroborate the information provided by tests.
3. Explain relevant measurement concepts as necessary for clarity at the level of detail that is appropriate for the intended audience(s).	3. Read the materials provided by test developers and avoid using tests for which unclear or incomplete information is provided.
4. Describe the process of test development. Explain how the content and skills to be tested were selected.	4. Become familiar with how and when the test was developed and tried out.

---

*Many of the statements in the Code refer to the selection of existing tests. However, in customized testing programs test developers are engaged to construct new tests. In those situations, the test development process should be designed to help ensure that the completed tests will be in compliance with the Code.

5. Provide evidence that the test meets its intended purpose(s).	5. Read independent evaluations of a test and possible alternative measures. Look for evidence required to support the claims of test developers.
6. Provide either representative samples or complete copies of test questions, directions, answer sheets, manuals, and score reports to qualified users.	6. Examine specimen sets, disclosed tests or samples of questions, directions, answer sheets, manuals, and score reports before selecting a test.
7. Indicate the nature of the evidence obtained concerning the appropriateness of each test for groups of different racial, ethnic, or linguistic backgrounds who are likely to be tested.	7. Ascertain whether the test content and norm group(s) or comparison group(s) are appropriate for the intended test takers.
8. Identify and publish any specialized skills needed to administer each test and to interpret scores correctly.	8. Select and use only those tests for which the skills needed to administer the test and interpret scores correctly are available.

B. Interpreting Scores

Test developers should help users interpret scores correctly.	Test users should interpret scores correctly.
Test Developers Should:	**Test Users Should:**
9. Provide timely and easily understood score reports that describe test performance clearly and accurately. Also explain the meaning and the process used to select the samples of test takers.	9. Obtain information about the scale used for reporting scores, the characteristics of any norms or comparison group(s), and the limitations of the scores.
10. Describe the population(s) represented by any norms or comparison group(s), the dates the data were gathered, and the process used to select the samples of test takers.	10. Interpret scores taking into account any major differences between the norms or comparison groups and the actual test takers. Also take into account any differences in test administration practices or familiarity with the specific questions in the test.
11. Warn users to avoid specific, reasonably anticipated misuses of test scores.	11. Avoid using test for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use.
12. Provide information that will help users follow reasonable procedures for setting passing scores when it is appropriate to use such scores with the test.	12. Explain how any passing scores were set and gather evidence to support the appropriateness of the scores.
13. Provide information that will help users gather evidence to show that the test is meeting its intended purpose(s).	13. Obtain evidence to help show that the test is meeting its intended purpose(s).

C. Striving for Fairness

Test developers should strive to make tests that are as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test Developers Should:

14. Review and revise test questions and related materials to avoid potentially insensitive content or language.
15. Investigate the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available. Enact procedures that help to ensure that differences in performance are related primarily to the skills under assessment rather than to irrelevant factors.
16. When feasible, make appropriately modified forms of tests or administration procedures available for test takers with handicapping conditions. Warn test users of potential problems in using standard norms with modified tests or administration procedures that result in non-comparable scores.

Test users should select tests that have been developed in ways that attempt to make them as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test Users Should:

14. Evaluate the procedures used by test developers to avoid potentially insensitive content or language.
15. Review the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available. Evaluate the extent to which performance differences may have been caused by inappropriate characteristics of the test.
16. When necessary and feasible, use appropriately modified forms of tests or administration procedures for test takers with handicapping conditions. Interpret standard norms with care in light of the modifications that were made.

D. Informing Test Takers

Under some circumstances, test developers have direct communication with test takers. Under other circumstances, test users communicate directly with test takers. Whichever group communicates directly with test takers should provide the information described below.

Test Developers or Test Users Should:

17. When a test is optional, provide test takers or their parents/guardians with information to help them judge whether the test should be taken, or if an available alternative to the test should be used.
18. Provide test takers the information they need to be familiar with the coverage of the test, the types of question formats, the directions, an appropriate test-taking strategies. Strive to make such information equally available to all test takers.

Under some circumstances, test developers have direct control of tests and test scores. Under other circumstances, test users have such control. Whichever group has direct control of tests and test scores should take the steps described below.

Test Developers or Test Users Should:

19. Provide test takers or their parents/guardians with information about rights test takers may have to obtain copies of tests and completed answer sheets, retake tests, have tests rescored, or cancel scores.

20. Tell test takers or their parents/guardians how long scores will be kept on file and indicate to whom and under what circumstances test scores will or will not be released.
21. Describe the procedures that test takers or their parents/guardians may use to register complaints and have problems resolved.

Note: The membership of the Working Group that developed the Code of Fair Testing Practices in Education and of the Joint Committee on Testing Practices that guided the Working Group was as follows:

Theodore P. Bartell
John R. Bergan
Esther E. Diamond
Richard P. Duran
Lorraine D. Eyde
Raymond D. Fowler
John J. Fremer
    (Co-chair, JCTP and Chair, Code Working Group)
Edmund W. Gordon
Jo-Ida C. Hansen
James B. Lingwall
George F. Madaus
    (Co-chair, JCTP)
Kevin L. Moreland
Jo-Ellen V. Perez
Robert J. Solomon
John T. Stewart
Carol Kehr Tittle
    (Co-chair, JCTP)
Nicholas A. Vacc
Michael J. Zieky
Debra Boltas and Wayne Camara
    of the American Psychological Association served as staff liaisons

Additional copies of the Code may be obtained from the National Council on Measurement in Education, 1230 Seventeenth Street, NW, Washington, D.C. 20036. Single copies are free.

Glossary of Assessment and Testing Terms

Accommodations See Testing Accommodations.

Accountability A system to evaluate whether, and to what degree, individuals meet standards or expectations. In educational contexts, accountability generally means holding teachers, schools, districts, and states responsible for student learning, as demonstrated by test scores and other assessments. However, accountability can also apply to students (e.g., promotion, graduation requirements) and education agencies (e.g., ensuring opportunity to learn, adequate funding).

Achievement Test An assessment that measures a student's current acquired knowledge and skills in one or more of the content areas common to most school curricula (e.g., reading, language arts, mathematics, science, social studies).

Administering Giving a test to a student or providing directions and support to a test taker.

Alternate Assessment An assessment used in place of a regular test because of the nature or the severity of a student's disability and the student's course of study. Generally, alternate assessments are designed to measure functional academic and literacy skills.

Alternative Assessment An assessment that differs from traditional achievement tests—for example, one that requires a student to generate or produce responses or products rather than answer only selected-response items. This type of assessment may include constructed-response activities, essays, portfolios, interviews, teacher observations, work samples, or group projects. See Authentic Assessment, Multiple Measures, Performance Assessment.

Analytic Scoring A scoring procedure in which a student's work is evaluated for selected characteristics, with each characteristic receiving a separate score.

Aptitude Test An assessment designed to predict a student's expected or potential acquisition of knowledge or skills.

Assessment The process of gathering information about a student's abilities or behavior for the purpose of drawing conclusions or making decisions about the

---

Note. Original source for the majority of information found in CTB McGraw-Hill. (1997). *Beyond the numbers: A guide to interpreting and using the results of standardized achievement tests.* Monterey, California: Author.

student's performance. Assessment comprises many tools or techniques (e.g., tests, observations, performance tasks), but is broader than any one of those tools or techniques.

Authentic Assessment An assessment that measures a student's performance in tasks and situations that resemble nonschool or real-life tasks. This type of assessment is closely aligned with and models what students do in the classroom. Examples include conducting science experiments, group problem-solving in mathematics, and writing for a public audience (e.g., a letter to the editor).

Bias The effect of a test when it systematically measures differently for different ethnic, cultural, regional, or gender groups.

Ceiling The upper limit of performance that can be measured effectively by a test. Individuals are said to have reached the ceiling of a test when they perform at the top of the range in which the test can make reliable discriminations. If an individual or group scores at the ceiling of a test, the test is too easy for him or her, and the next higher level of the test, if available, should be administered.

CIA Alignment Stands for the coordination and common content covered in the curriculum, instruction, and assessment.

Construct Validity A test is said to have construct validity when the particular knowledge domain or behavior reported to be measured is actually measured.

Constructed-Response Item An assessment unit with directions and a question or a problem that elicits a written, pictorial, or graphic response. Sometimes called an *open-ended item.*

Content Standards A statement or description of the knowledge and skills in a content area (e.g., language arts, mathematics, science, social studies) that students should learn (and that teachers should teach). National content standards are published by groups such as the National Council of Teachers of Mathematics and the International Reading Association/National Council of Teachers of English; local content standards are produced by states and/or school districts. Content standards are sometimes called *standards* or *academic standards.*

Criterion-Referenced Test (CRT) An assessment that measures a student's performance according to specified standards or criteria rather than in comparison to the performances of other test takers.

Criterion-Related (Referenced) Score Technically a criterion-related test score is one that is interpreted by comparing a person's performance to a known set of standards or model of performance. This type of score interpretation is often contrasted with a norm-referenced score, in which meaning is gained by comparing one person to a group of people used to norm the test.

Curriculum-Referenced Test An assessment that measures what a student knows or can do in relation to specific, commonly taught curriculum objectives.

Derived Score Any score that is based on (derived from) a scale score. Examples are national percentile scores, normal curve equivalent scores, and grade-equivalent scores.

Distracter (or Distractor) An incorrect answer choice in a selected-response or matching test item. Sometimes called a *foil.*

Equated Score A score from one test that is equivalent to a score from another test. Equated scores are usually obtained by administering the two tests of interest to a representative sample of students. Scores from one test are then aligned with scores on the other test using equating analyses.

Equivalent Form Any of two or more forms of a test, usually standardized on the same population and published at the same time; designed to be similar in item content and difficulty so that scores on the forms will be comparable.

Floor The opposite of ceiling, the floor is the lower limit of performance that can be measured effectively by a test. Individuals are said to be at the floor of a test when they perform at the bottom of the range in which the test can make reliable discriminations. If an individual or group scores at the floor of a test, the test is too difficult for that individual or group, and the next lower level of the test, if available, should be administered.

Frequency Distribution An ordered tabulation of individual scores (or groups of scores) that shows the number of persons who obtained each score or placed within each group of scores.

Functional Range The functional range of a test is the range of grades for which the test can be administered in order to obtain accurate norm-referenced data. For most tests, this range is two grades above or below the grade for which the test was intended.

Grade Equivalent (GE) A score on a scale developed to indicate the school grade (usually measured in 10ths of a year) that corresponds to an average test score. A grade equivalent of 6.4 is interpreted as a score that is average for a group that has completed the fourth month of Grade 6. Grade equivalents do not compose a scale of equal intervals and are not usable in drawing profiles.

Grade Mean Equivalent A derived score expressed as the grade placement of those students for whom a given score was average.

Holistic Scoring A scoring procedure yielding a single score based on overall student performance rather than an accumulation of points. Holistic scoring uses rubrics.

Interpreting Translating a score so that it has meaning; this often involves the use of criterion-referenced scores or norm-referenced scores.

Item One of the assessment units, usually a problem or a question, that is included on a test.

Item Bias The effect of an item when it systematically measures differently for different ethnic, cultural, regional, or gender groups.

Large-Scale Assessment An approach to testing whereby an entire population of students (e.g., all fourth-graders, all eighth-graders) are administered an achievement test as part of an accountability system.

Levels of Performance An approach to interpreting results on a test that translates scores within various ranges by using descriptions of performances that communicate a continuum of proficiency. Examples include Minimal proficiency, Basic proficiency, Proficient, and Advanced proficiency.

LEP Stands for limited English proficiency.

Mean A measure of central tendency. An average calculated by adding a set of scores and dividing by the number of scores in the set.

Measure To quantify or to place a number on a student's performance.

Median A measure of central tendency. The middle score in a set of ranked scores.

Mode A measure of central tendency. The most frequently obtained score.

Modifications Changes that alter the level or content of the test. Examples include giving a lower grade level of a test or deleting or changing the content of a test. Modifications are distinct from testing accommodations in that testing accommodations change noncontent aspects of a test.

Multiple Choice See Selected-Response Item

Multiple Measures Assessments that measure student performance in a variety of ways. Multiple measures may include standardized tests, teacher observations, classroom performance assessments, and portfolios.

National Percentile Same as percentile, but based on a national norm group (see Percentile).

Nonstandard Accommodation An accommodation that is not generally approved or endorsed, either due to lack of previous research, or because of educational policies guiding accommodation use.

Normal Curve Equivalent (NCE) An equal-interval scale score that can be treated arithmetically; a normalized standard score with a mean of 50 and a standard deviation of 21.06.

Norm-Referenced Test (NRT) A standardized assessment—that is, an assessment in which all students perform under the same conditions—that compares a student or group of students with a specified reference group, usually others of the same grade or age. Technically, tests are not norm referenced; rather, the scores a test produces are norm referenced (e.g., percentiles, NCEs).

Number-Correct or "Raw" Score The number of correct responses (NCR) is the number of items answered correctly by a student on any given test section.

Open-Ended Item See Constructed-Response Item.

Percentage Correct The proportion of items correct divided by the total number of items.

Percentile (PR) The most frequently used score for describing achievement test results to persons outside the test and measurement community; percentiles are essentially "counting scores" that designate the proportion of students in the norming sample for a given grade whose scores fell at or below a certain point. For example, a score representing the 26th percentile is a score that is equal to or better than 26% of the scores in the distribution. Percentiles are not at equal intervals (e.g., the gap between the 1st and 2nd percentiles is roughly equal to the gap between the 38th and 50th percentiles). Often confused with percentage of items correct.

Percentile Rank Same as percentile.

Performance Assessment An assessment activity that requires students to construct a response, create a product, or perform a demonstration. Usually there are multiple ways to approach a performance assessment, and more than one correct answer. Performance assessments are often, but not always, aligned to performance and content standards.

Performance Standards An objective statement of what students will do to demonstrate mastery of content standards. Typically, performance standards are fixed to grade levels, so that students in earlier grades are expected to produce different performances than students in later grades to mastery of the standards. However, some use the terms *performance standards* and *proficiency standards* interchangeably. See Content Standards and Proficiency Standards.

Portfolio Assessment An assessment based on a collection of evidence highlighting a student's performance, often over a period of time. A portfolio assessment can contain a wide variety of information, including norm-referenced test scores, awards, drawings, audio or video tapes, and writing samples. Sometimes these are best examples, and sometimes they are a representative sample, exhibiting a record of changes in performance.

Primary Trait Scoring A scoring procedure in which a student's work is evaluated for one or more specific traits or dimensions; other traits or dimensions are not scored. For example, a student's writing may be scored on organization, but not on grammar or spelling. Also known as *trait scoring*.

Proficiency Standards An objective statement of level of performance required for students to demonstrate proficiency in a standard. Most often, performance standards are scores on tests (e.g., scale scores) or other assessments (e.g., proficiency ratings of performance assessment or portfolios). Proficiency standards differ from performance and content standards in that content standards define what students should know, performance standards define what students should do, and proficiency standards define how well students must do to demonstrate mastery of content standards.

Prompt An assessment topic, situation, or statement to which students are expected to respond. See also Stimulus.

Reliability The degree to which an assessment yields consistent results over different items/tests, times, settings, or raters. A test is said to be reliable to the extent that a student's scores are nearly the same across different items, settings, times, or judges. Consistency across items or parallel tests is termed *internal consistency*. Consistency across settings and times is termed *test–retest reliability* or *stability*. Consistency across judges is termed *interrater reliability*. The key characteristic of a reliable test is consistency. Reliability is a necessary, but not sufficient condition for validity.

Rubric A scoring tool, or set of criteria, used to evaluate a student's test or performance. Rubrics may be diagnostic or analytic, in that they specify ratings for multiple characteristics of student work, or rubrics may be holistic, in that they describe a single, global rating for student work. Rubrics are often shared with students to help them understand task demands, and they may be completed by students and teachers to evaluate student work.

Scale An organized set of measurements, all of which measure one property or characteristic and imply at least an ordinal ranking (e.g., from low to high). Different types of test-score scales use different units (e.g., number correct, percentiles, and IRT scale score). Scales provide a metric for measurement (e.g., Celcius and Fahrenheit are different scales for measuring temperature).

Scale Score Loosely defined as any derived score; more technically, any of several systems of scores used to compare scores across different forms of the same test.

Selected-Response Item A question or incomplete statement that is followed by answer choices, one of which is the correct or best answer. Also referred to as a *multiple-choice* item.

Standard Score A general term referring to scores that have been "transformed" for reasons of convenience, comparability, or ease of interpretation and fixed to the center of a normative sample. All standard scores are defined by the distance from the mean of the normative sample and the spread of scores from the center (standard deviation). For example, z scores have a mean of 0 and a standard deviation of 1; NCEs have a mean of 50 and a standard deviation of 21.06; most individual tests yield standard scores with a mean of 100 and a standard deviation of 15. Standard scores can be translated from one standard score scale to another, just as it is possible to translate kilometers to miles or kilograms to pounds.

Standard Deviation Approximately the average distance of individual scores from the mean. In a normal distribution, 68% of test scores are within 1 standard deviation of the mean. About 95% of scores fall within 2 standard deviations of the mean.

Standard Error of Measurement (SEM) Defines a range within which a student's "true score" would be likely to fall had that student taken the test numerous times. All tests have inherent measurement error because they are a sample of student performance at one particular time.

Standardized Test An assessment with directions, time limits, materials, and scoring procedures designed to remain constant each time the test is given, to ensure comparability of scores. Many standardized tests have norms. All norm-referenced tests are standardized.

Stanines A scale that divides scores of the norm population into 9 equidistant groups. Similar to NCEs, but because they range from only 1 to 9, they provide a less precise description of the score's position in the normal distribution than an NCE. The proportion of scores at each stanine is 4, 7, 12, 17, 20, 17, 12, 7, and 4, respectively. The first three stanines are often interpreted as being "below average," the next three as "average," and the top three as "above average."

Stem The part of an item that asks a question, provides directions, or presents a statement to be completed.

Stimulus A passage or graphic display about which questions are asked.

Test A device or procedure designed to elicit responses that permit an inference about what the test taker knows or can do.

Test Battery A set of several tests designed to be given as a unit.

Test Objective A targeted goal that can be measured by an assessment.

Test Stimulus See Stimulus.

Testing Accommodations Changes in the way a test is administered to a student or responded to by a student. Testing accommodations are intended to offset distortions in test scores caused by a disability without invalidating or changing what the test measures. Common testing accommodations involve extra time, assistance with directions, assistance with reading, and enlarged print size.

Tests A structured method or procedure through which educators obtain evidence about a student's ability or behavior.

Usability A characteristic of a test that is practical, not technical. It generally concerns the ease of use, the skills required to administer it, the time needed to score it, and the meaningfulness of the scores—that is, the usefulness of the test.

Validity The degree to which an assessment measures what it is intended to measure. Validity defines how assessment results should, and should not, be interpreted with regard to a particular use or purpose.

References

Airasian, P. W. (1994). *Classroom assessment* (2nd ed.). Boston: McGraw-Hill.

American Educational Research Association. (1988). *Code of Fair Testing Practices in Education.* Washington, DC: Author.

American Educational Research Association. (1999). *Standards for Educational and Psychological Testing.* Washington, DC: Author.

American Educational Research Association. (2000). *AERA position statement concerning high-stakes testing in preK–12 education.* Washington, DC: Author. Available: http://www.aera.net/about/policy/stakes.htm

American Federation of Teachers. (1990). *Standards for Teacher Competence in Educational Assessment of Students.* Washington, DC: Author.

Barton, P. E. (1999). *Too much testing of the wrong kind; too little of the right kind in K–12 education. A policy information perspective.* Princeton, NJ: Educational Testing Service, Policy Information Center.

Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K. A., & Valadez, J. R. (1993). Mathematics performance assessment: Technical quality and diverse student impact. *Journal for Mathematics Education, 24,* 190–216.

Bielinski, J., Thurlow, M., Minnema, J., & Scott, J. (2000). *How out-of-level testing affects the psychometric quality of test scores* (Out-of-Level Testing Rep. No. 2). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Canner, J., et al. (1991). *Regaining trust: Enhancing the credibility of school testing programs* (Mimeo). National Council on Measurement in Education Task Force.

Cotton, K. (1999). *Research you can use to improve results.* Alexandria, VA: Association for Supervision and Curriculum Development.

Council for Exceptional Children. (1998). *What every special educator must know: The International Standards for the Preparation and Licensure of Special Educators* (3rd ed.). Arlington, VA: Author.

CTB/McGraw-Hill. (1996). *Teacher's guide to TerraNova.* Monterey, CA: Author.

Demaray, M. K., & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13,* 8–24.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives, 6*(1). Available: http://olam.ed.asu.edu/epaa/v6n1. html

Education Week. (2001). Quality counts 2001: A better balance. *Editorial Projects in Education, 20*(1). Available: http://www.edweek.org/stereports/qc01

Elliott, S. N., & Braden, J. P. (2000). *Educational assessment and accountability of all students.* Madison: Wisconsin Department of Public Instruction.

Elliott, D. N., & Kratochwill, T. R. (1996). *Experimental analysis of the effects of testing accommodations on the scores of students with disabilities* (U.S. Department of Education Grant No. 84.023). Madison: University of Wisconsin-Madison, Wisconsin Center for Education Research.

Elliott, S. N., & Kratochwill, T. R. (1998–2001). *Experimental analysis of the effects of testing accommodations on the scores of students with disabilities* (U.S. Department of Education Grant No. 84.023). Madison: University of Wisconsin-Madison, Wisconsin Center for Education Research. **[Au: not cited in text. Cite or delete.]**

Elliott, S. N., Kratochwill, T. R., & McKevitt, B. C. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities. *Journal of School Pyschology, 39*(1), 3–24.

Elliott, S. N., Kratochwill, T. R, & Schulte, A. G. (1999). *The assessment accommodations guide.* Monterey, CA: CTB/McGraw-Hill.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C., & Karns, K. (2000). Supplementing teacher judgments of test accommodations with objective data sources. *School Psychology Review, 29*, 65–85.

Glidden, H. (1998). *Making standards matter 1998: An annual fifty-state report on efforts to raise academic standards.* Washington DC: American Federation of Teachers.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice, 17*(2), 16–19.

Gresham, F. M., Reschly, D. J., & Carey, M. (1987). Teachers as "tests": Classification accuracy and concurrent validation in the identification of learning disabled children. *School Psychology Review, 16*, 543–553.

Haertel, E. H. (1999). Validity arguments for high stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.

Hammer, D. (1998). *The standards teacher: Standards for excellence in education.* Alexandria, VA: Association for Supervision and Curriculum Development.

Hanley, W. M., Madaus, G. F., & Lyons, G. (1993). *The fractured marketplace for standardized testing.* Boston: Kluwer.

Heubert, J. P., & Hauser, R. M. (Eds.). (1998). *High stakes: Testing for tracking, promotion, and graduation.* Washington, DC: National Academy Press.

Hoge, R. D., & Coladarci, T. (1989). Teacher based judgments of academic achievement: A review of the literature. *Review of Educational Research, 59*, 297–313.

Huemann, J. E., & Warlick, K. R. (2000). *Memorandum: Questions and answers about provisions in the Individuals with Disabilities Education Act Amendments of 1997 to students with disabilities and state and district-wide assessments.*

Kean, M. (1998). *Education assessment: A primer for school boards.* Monterey, CA: CTB/McGraw-Hill.

Kleinert, H. L., Haig, J., Kearns, J. F., & Kennedy, S. (2000). Alternate assessments: Lessons learned and roads to be taken. *Exceptional Children, 67,* 51–66.

Koretz, D. (1997). *The assessment of students wtih disabilities in Kentucky.* (CSE Tech. Rep. No. 431). Los Angeles: Center for Research on Standards and Student Testing.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–14. Available: http://www.aera.net/pubs/er/arts/29–02/linn01.htm

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

McDonnell, L. M., McLaughlin, M. J., & Morison, P. (Eds.). (1997). *Educating one and all: Students with disabilities and standards-based reform.* Washington, DC: National Academy Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 144–188). New York: Macmillan.

Meyer, R. L. (1996). Value-added indicators of school performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 171–196). Washington, DC: National Academy Press.

Minnema, J., Thurlow, M., Bielinski, J., & Scott, J. (2000). *Past and present understandings of out-of-level testing: A research synthesis* (Out-of-Level Testing Rep. No. 1). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

National Center on Educational Outcomes. (1999). *Participation of students with disabilities.* Minneapolis: Author. Available: http://wwww.coled.umn.edu/NCEO/TopicAreas/Participation/participation_FAQ.htm

National Center on Educational Outcomes. (2000, May). Non-approved accommodations: Recommendations for use and reporting. *NCEO Policy Directions* (No. 11). Minneapolis: Author.

Newmann, F. M., Marks, H. M., & Gamoran, A. (1995). *Authentic pedagogy and student performance.* Madison: Wisconsin Center for Education Research.

Newmann, F. M., & Wehlage, G. G. (1995). *Successful school restructuring: A report to the public and educators by the Center on Organization and Restructuring of Schools.* Madison: Wisconsin Center for Education Research.

Novello, M. K. (1999, March). *Recent research in math, science, language arts, social studies, and the arts.* Paper presented at the annual meeting of the Association for Supervision and Curriculum Development, San Francisco, CA. (ERIC Document Reproduction Service No. ED 433 244).

Phelps, R. P. (1996). Test basher benefit-cost analysis. *Network News and Views, 15*(3), 1–16.

Phillips, S. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7,* 93–120.

Samson, G. E. (1985). Effects of training in test-taking skills on achievement test performance: A quantitative synthesis. *Journal of Educational Research, 78*(5), 261–265.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive domain. *Review of Educational Research, 49,* 60–79.

Schulte, A. G. (2000). *Experimental analysis of the effects of testing accommodations on students' standardized mathematics test scores.* Unpublished doctoral dissertation, University of Wisconsin-Madison.

Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal, 31,* 231–262.

Thurlow, M. L., Elliott, J. L., & Ysseldyke, J. E. (1998). *Testing students with disabilities: Practical strategies for complying with district and state requirements.* Thousand Oaks, CA: Corwin.

Thurlow, M. L., House, A., Boys, C., Scott, D., & Ysseldyke, J. (2000). *State participation and accommodation policies for students with disabilities: 1999 update* (Synthesis Rep. No. 33). Minneapolis: National Center on Educational Outcomes.

Thurlow, M. L., Nelson, J. R., Teelucksingh, E., & Ysseldyke, J. E. (2000). *Where's Waldo? A third search for students with disabilities in state accountability reports* (Tech. Rep. No. 25). Minneapolis: National Center on Educational Outcomes.

Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1993). *Testing accommodations for students with disabilities: A review of the literature.* (Synthesis Rep. No. 4). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Tindal, G., & Fuchs, L. (1999). *A summary of research on test accommodations: What we know so far.* (Tech. Rep.). Lexington: University of Kentucky, Mid-South Regional Resource Center.

Tindal, G., Glasgow, A., Helwig, B., Hollenbeck, K., & Heath, B. (1998). *Accommodation in large scale tests for students with disabilities: An investigation of reading math tests using video technology.* Unpublished manuscript, Council of Chief State School Officers, Washington, DC.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64,* 439–450.

Trimble, S. (1998). *Performance trends and use of accommodations on a statewide assessment: Students with disabilities in the KIRIS on-demand assessments from 1992–93 through 1995–96.* (State Assessment Series, Maryland/Kentucky Rep.). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Turner, M. D., Baldwin, L., Kleinert, H. L., & Kearns, J. F. (2000). The relation of a statewide alternate assessment for students with severe disabilities to other measures of instructional effectiveness. *The Journal of Special Education, 34*(2), 69–76.

Webb, N. L. (1997). *Determining alignment of expectations and assessments in mathematics and science education* (Vol. 1, No. 2). Madison: National Institute for Science Education. Available: http://www.wcer.wisc.edu/NISE/Publications/Briefs/Vol_1_No_2/

Witt, J. C., Elliott, S. N., Daly, E. J. III, Gresham, F. M., & Kramer, J. J. (1998). *Assessment of at-risk and special needs children* (2nd ed.). Boston: McGraw-Hill.

Ysseldyke, J. E., & Olsen, K. (1999). Putting alternate assessments into practice: What to measure and possible sources of data. *Exceptional Children, 65,* 175–186.

Ysseldyke, J. E., Thurlow, M., Erickson, R. Gabrys, R., Haigh, J., Trimble, S., & Gong, B. (1996). *A comparison of state assessment systems in Kentucky and Maryland with a focus on the participation of students with disabilities.* (State Assessment Series, Maryland/Kentucky Rep. No. 1). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Council for Exceptional Children

9 780865 863750

BEST COPY AVAILABLE